Spatio-temporal predictions of the abundance of biting midges (Culicoides) in the western part of Europe: A large scale machine learning approach based on surveillance data, remote sensing and climate predictors

Cuellar, Ana Carolina

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Spatio-temporal predictions of the abundance of biting midges (*Culicoides*) in the western part of Europe:

A large scale machine learning approach based on surveillance data, remote sensing and climatic predictors

PhD Thesis

Ana Carolina Cuéllar

Lyngby, November 2018

Division for Diagnostics & Scientific Advice,
National Veterinary Institute
Technical University of Denmark, Lyngby, Denmark

Supervisors

Epidemiologist René Bødker (Main supervisor)
Division for Diagnostics & Scientific Advice – Epidemiology
National Veterinary Institute - Technical University of Denmark
Denmark

Postdoc Lene Jung Kjær (Co-supervisor)
Division for Diagnostics & Scientific Advice – Epidemiology
National Veterinary Institute - Technical University of Denmark
Denmark

Professor Nils Toft (Co-supervisor)
Division for Diagnostics & Scientific Advice – Epidemiology
National Veterinary Institute - Technical University of Denmark
Denmark

Assessment Committee

Associate Professor Lasse Engbo Christiansen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Denmark

Senior Scientist Dr Beth V. Purse
Center for Ecology & Hydrology
UK

Professor Flemming Konradsen
School of Global Health
University of Copenhagen
Denmark

Preface & Acknowledgements

The readers should know that this thesis started at least two years before my first day at DTU vet in 2015. It started with Rene Bødker (my supervisor) and Carsten Kirkeby (postdoc back then) contacting the main researchers from different European countries. They were in charge of creating the biggest data set gathered for *Culicoides* until date in Europe!. This was a hard task that, I imagine, requires a lot of dedication and patience. So one the of the most difficult part of this thesis (data collection), was in charge of this two guys. Thanks for that!.

I arrived to DTU and the data was waiting for me, a bit pre-processed by Wesley Tack (a nice biologist working at that time in AVIA GIS) who merged the data collected and put it together into a single dataset and it was this dataset that, after three years of coexistence together, gave me sometimes headaches but also happy moments (like the one right now, while I write preface at the end of this PhD). This PhD was quite intensive and interesting journey, at least in the last part, and I am so happy that I had the chance to do it. I am surprised of how much I learnt during these years (well done me) and I hope to continue learning more.

First of all I would like to thank to my supervisor Rene Bødker for giving some Argentinean girl the chance to do a PhD. Also for all the shared knowledge and the interesting conversations regarding *Culicoides* and statistics and modelling and all kind of fun topics. Thanks a lot for the support I got in the tuff moments (including tears) during this PhD. Special thanks to Lene Jung Kjær for all the support during these years and caring for me in the difficult times. I thank also Nils Toft, for his support as co supervisor.

I would also like to thank to all my co-authors : Andreas Baum, Anders Stockmarr, Henrik Skovgard, Søren Achim Nielsen, Gunnar Anderson, Anders Lindström, Jan Chirico, Renke Lühken, Sonja Steinke, Ellen Kiel, Jörn Gethmann, Franz J. Conraths, Magdalena Larska, Marcin Smreczak, Anna Orłowska, Inger Hamnes, Ståle Sviland, Petter Hopp, Katharina Brugger, Franz Rubel, Thomas Balenghien, Claire Garros, Ignace Rakotoarivony, Xavier Allène, Jonathan Lhoir, David Chavernac, Jean-Claude Delécolle, Bruno Mathieu, Delphine Delécolle, Marie-Laure Setier-Rio, Roger Venail, Bethsabée Scheid, Miguel Ángel Miranda Chueca, Carlos Barceló, Javier Lucientes, Rosa Estrada, Alexander Mathis

and Wesley Tack. Thanks for sharing the data and especially for the valuable comments regarding the work carried on during this PhD. Special thanks to Anders Stockmarr and Andreas Baum for their valuable contribution to our modelling papers and incredible and valuable scientific support during the last year of my PhD.

Thanks to Guy Hendrickx, Els Ducheyne and Wesley Tack from the private company AVIA-GIS, in Belgium, who received me for two weeks, when I was a junior PhD, in their offices and they gave me the initial push in the modelling world.

I would like to thank Gunnar Andersson for his supervision during my external research in the National Veterinary Institute (SVA) in Uppsala. Thanks Gunnar for the nice Random Forest discussions, I learned a lot from you, and also thank you for taking me sailing from Stockholm to Uppsala, a remarkable experience I will never forget. Anders Lindström, also from SVA in Uppsala, thank you so much for making me feel so welcome in SVA and for open the doors of your house to me (dinner included), I hope our paths will cross again.

Thanks to all my friends and colleagues from the Epi group at DTU. It has been a pleasure to work with you guys. Maya, thank you so much so the help with R, with all the "Maya, could you help me with this?". If it wasn't for you I might still be solving problems in R. Special thanks to my Bangladeshi friend Najmul Haider who, despite being a fan of the Brazilian football team, I admire a lot for being a great scientist and a great person. Thanks to Ana Carolina for all the brunch and hang out meetings.

Thanks to Niels, my beloved boyfriend, for his love and support during the last of my PhD (he had to deal with the worst time of it!). Thanks for cooking for me when I was very busy (an when I was not) and for being the sweetest partner in the world.

Thanks to my not DTU friends in Copenhagen : Gachi, Eduardo and Mikael. Thanks for the nice moments spent together, and for being such a good friends, always there for me. And also my old friends in Argentina: Claudia, Flor, Ale, my ex colleagues and friends from my master program (you have no idea how much I love you and miss guys!).

10.000.000 thanks to Fernanda García, who gave me a lot of support especially in this last part of the PhD and took charge of designing this beautiful thesis (she worked until 9 pm in the evening for me) Thanks a

lot my dear friend!.

Special thanks to my best friend Romina Diaz Gomez who, even the nine hours difference that separate us (she lives in California), supported me during all this time and made me feel that distance is not important for keeping a tight friendship. Te quiero mucho changa!

And last but not least thanks to my beloved family in Argentina: Mom (Estela), Dad (Juan), siblings (Mariví, Juan Pablo and Matías), niece (Luchi) and nephews (Gianfranco, Lucas and Tomy) for their support. I wouldn't be where I am not if it wasn't for them. I love you.

And thanks to all the people who reads this thesis, I hope you find it interesting and more important, enjoyable!

Ana Carolina Cuéllar

Summary

In the veterinary field, *Culicoides*-borne diseases cause considerable economic losses as they affect animal welfare, reduce animal production and stop animal trade between infected countries and those free of disease. Other indirect costs are related to surveillance and vaccination programs.

Bluetongue virus (BTV) and Schmallenberg virus (SBV) are transmitted to ruminant livestock (cattle, goats and sheep) by the bite of female *Culicoides* "biting midges". Bluetongue disease (BT) was reported in southern Europe during the 19th century, where bluetongue appeared in some countries of the Mediterranean Basin. On these occasions, the disease occurrence corresponded well with the known distribution of the afro-tropical species *C. imicola.* Bluetongue has never been reported in Northern Europe until 2006, when an unexpected BTV-8 strain outbreak started. The disease spread and affected several countries and caused a huge economic loss to the European Union. During this outbreak, surveillance programs started in many of the affected countries and the results showed that *C. imicola* was completely absent in the area. Species of the Obsoletus and Pulicaris ensemble were then suspected of transmitting the virus, and this was confirmed when the BTV – 8 was isolated from wild specimens caught in the field. Schmallenberg disease (SB) was discovered in Germany in 2011 and spread among 22 European countries. Outbreaks of this disease caused congenital malformations and stillbirth in cattle, sheep, and goats resulting in economic losses.

Since the start of the BT epidemic, European Union member states started to carry out entomological surveillance programs in order to determine the vector species composition and seasonal dynamics. Many countries have published the results from the national surveillance. Some of them presented distribution maps showing the distribution of the vectors, at a national scale. Nevertheless, there is still a need to generate *Culicoides* vector distribution and abundance maps at a continental level, as European legislation for vector-borne diseases is founded on joint decisions among the member states. Therefore, entomological data from European countries were gathered by the VICE EMIDA project (Vector-borne Infections: risk based and cost effective surveillance systems) in order to create a large transnational *Culicoides*

dataset for Europe. Nine countries agreed to share data: Spain, France, Germany, Austria, Switzerland, Denmark, Sweden, Norway and Poland.

The aim of this thesis was to analyse the available *Culicoides* data collected on European farms to understand the spatial and temporal dynamics of the main vectors of BT and SB in Europe at a continental scale. We gathered existing entomological data collected in farms from nine European countries and predicted the occurrence and abundance of these vectors for large areas in Europe, from southern Spain to northern Sweden. All analyses were carried out individually for the Obsoletus and Pulicaris ensembles and for *C. imicola*.

In a first step, we carried out a descriptive analysis of the entomological data at a continental level and explored the temporal fluctuation of mean abundance summarized by latitudinal ranges of 5° width. We also mapped the monthly mean abundance in the sampled farms and used interpolation in order to detect spatial trends in the observed abundance. Additionally, we estimated the start of the season at NUTS level to highlight any temporal trend. The main findings of this analysis were that the Obsoletus ensemble, the main vector of BT in the Palearctic Europe, showed a south-north trend in which the highest cumulative vector abundance was found towards the north, even though the length of the vector season was shorter in the north. This might be the result of an adaptation of the Obsoletus ensemble to colder areas.

In a second step, we modelled Presence/Absence data using the machine learning technique Random Forest (RF) based on climatic data and environmental data from remote sensing. We produced maps of the vector probability of presence per month. These maps were classified into three classes according to two thresholds calculated to minimize class misclassification. The three classes were: Presence, Absence and an intermediate class representing areas with an uncertain status where surveillance efforts should be in focus. RF performed well for the Obsoletus ensemble, fair for the Pulicaris ensemble and very well for *C. imicola*. The maps presented are useful as a base tool for decision making regarding restrictions on animal movement or for implementation of surveillance programs in those areas were the model was not able to classify Presence or Absence classes.

Using the same modelling approach with climatic data and environmental data from remote sensing, we modelled the abundance

by month and created monthly abundance maps for the Obsoletus and Pulicaris ensembles and *C. imicola*. RF performed well for predicting the abundance for the Obsoletus ensemble, fair for the Pulicaris ensemble and poor for *C. imicola*, even though the geographic distribution coincided with the known distribution for this species. When predictions are fair, the predicted values might be useful for the generation of risk assessment models, for instance R_0 models.

Lastly, we compared RF performance to simple interpolation performance (IDW algorithm) and the results showed that interpolation performed only slightly poorer than RF. This highlight the limitations for modelling *Culicoides* abundance data based on environmental and climatic predictors and in the analysis in this thesis, several biases were introduced as a consequence of merging data from different countries using different protocols.

The findings of this thesis serve to understand the spatio-temporal dynamics of the main vector of BT and SB in Europe, allowing for improved decisions for disease control and surveillance programs. Knowing vector spatial distribution is an important step delimitating the boundaries of a possible outbreak of vector-borne diseases, as the diseases cannot spread beyond the areas where the vector is present. Therefore, mapping the vector distribution and abundance can be used as a fundamental input for more general models that, together with information of other biological parameters, could aid in creating R_0 maps for *Culicoides* borne infections, useful for improving the response to new disease outbreaks.

Resumé

Inden for veterinærområdet forårsager*Culicoides*-bårne sygdomme betydelige økonomiske tab, da de påvirker dyrevelfærd, reducerer animalsk produktion og stopper handel og transport med dyr mellem inficerede og sygdomsfrie lande. Andre indirekte omkostninger relaterer sig til overvågnings- og vaccinationsprogrammer.

Bluetongue virus (BTV) og Schmallenberg virus (SBV) overføres til drøvtyggere (kvæg, geder og får) ved bid af hunnerne af *Culicoides* mitter. Bluetongue (BT) blev rapporteret i Sydeuropa i det 20. århundrede, hvor sygdommen dukkede op i nogle lande i Middelhavsområdet. Her korrelerede sygdomstilfældene med den kendte fordeling af den afro-tropiske art *C. imicola*. BT er aldrig blevet rapporteret i Nordeuropa før i 2006, da et uventet BTV-8-udbrud startede. Sygdommen spredte sig og ramte flere lande og forårsagede store økonomiske tab i EU. Under dette sygdomsudbrud startede man overvågningsprogrammer i mange af de berørte lande, og resultaterne viste, at *C. imicola* var fuldstændig fraværende i området. Arter fra Obsoletus- og Pulicaris-ensemblet blev derefter mistænkt for at overføre virusset, og dette blev bekræftet, da BTV-8 blev isoleret fra disse ensembler i felten. Schmallenberg (SB) blev opdaget i Tyskland i 2011 og spredte sig til 22 europæiske lande. Udbrud af denne sygdom forårsager medfødte misdannelser og dødfødsler hos kvæg, får og geder, hvilket resulterede i økonomiske tab.

Siden starten af BT-epidemien begyndte EU-medlemsstaterne at gennemføre entomologiske overvågningsprogrammer med henblik på at bestemme sammensætningen af vektorarter og deres sæsondynamik. Mange lande har publiceret resultaterne fra den nationale overvågning. Nogle af landene har udviklet distributionskort, der viser fordelingen af vektorerne på nationalt plan. Ikke desto mindre er der stadig behov for at generere *Culicoides* vektor distributions- og vektor densitetskort på et kontinentalt plan, da europæisk lovgivning for vektorbårne sygdomme er baseret på fælles beslutninger blandt medlemsstaterne. Derfor blev entomologiske data fra flere europæiske lande samlet af VICE EMIDA-projektet (Vector-borne Infections: risk based and cost effective surveillance systems) for at skabe et stort transnationalt *Culicoides* datasæt for Europa. Ni lande blev enige om at dele data: Spanien, Frankrig, Tyskland, Østrig, Schweiz, Danmark, Sverige, Norge og Polen.

Formålet med denne afhandling var at analysere de tilgængelige *Culicoides* data indsamlet på europæiske bedrifter for at forstå den rumlige og tidsmæssige dynamik af de vigtigste vektorer af BT og SB i Europa på et kontinentalt plan. Vi samlede eksisterende entomologiske data fra gårde fra ni europæiske lande og prædikterede forekomst og densitet af disse vektorer til store områder i Europa, fra det sydlige Spanien til det nordlige Sverige. Alle analyser blev udført individuelt for Obsoletus- og Pulicaris-ensemblerne og for *C. imicola*.

Som det første skridt, lavede vi en deskriptiv analyse af det entomologiske data på et kontinentalt plan og undersøgte temporære fluktuationer af gennemsnitlig tæthed opsummeret per 5° bredde. Vi kortlagde også den månedlige gennemsnitlige densitet i de gårde, data kom fra, og anvendte interpolering for at undersøge mulige rumlige tendenser i den observerede tæthed. Derudover estimerede vi starten af sæsonen på NUTS-niveau for at identificere tidsmæssige tendenser. De vigtigste resultater af denne analyse var, at Obsoletus-ensemblet, den vigtigste vektor af BT i det Palearktiske Europa, viste en syd-nord-trend, hvor den højeste kumulative vektordensitet blev fundet i de nordlige egne, selv om længden af vektorsæsonen var kortere i disse nordlige egne. Dette kan være resultatet af, at arterne i Obsoletus-ensemblet har tilpasset sig koldere områder.

Som næste skridt modellerede vi udbredelsen af mitter ved hjælp af machine learning-teknikken Random Forest (RF) baseret på klimatiske data og miljødata fra remote sensing. Vi producerede kort, der angav sandsynligheden for vektorforekomsten per måned. Disse kort blev klassificeret i tre klasser i forhold til to tærskelværdier beregnet for at minimere misklassificering. De tre klasser var: Tilstedeværelse, Fravær og en mellemklasse, der repræsenterer områder med en usikker status, hvor man bør fokusere overvågningsindsatsen. RF-metoden fungerede godt for Obsoletus-ensemblet, nogenlunde for Pulicaris-ensemblet og meget godt for *C. imicola*. De præsenterede kort er nyttige som et grundlæggende redskab for beslutningstagning i forhold til restriktioner af dyretransporter eller for etablering af overvågningsprogrammer i de områder, hvor modellen ikke kunne klassificere tilstedeværelses- eller fraværsklasser.

Ved at bruge den samme modelleringsmetode med klimatiske data og miljødata fra remote sensing, modellerede vi tæthed per

måned og producerede månedlige tæthedskort for Obsoletus- og Pulicaris-ensemblerne og *C. imicola*. RF fungerede godt til at forudsige tæthed af Obsoletus-ensemblet, nogenlunde for Pulicaris-ensemblet og dårligt for *C. imicola*, selvom den geografiske fordeling stemte overens med den kendte fordeling for denne art. Når prædiktionerne er rimelige, kan de prædikterede densitetssværdier være nyttige til brug i risikovurderingsmodeller, f.eks. R_0-modeller.

Endelig sammenlignede vi RF metoden med en simpel interpolation (IDW-algoritme), og resultaterne viste, at interpolering kun var en smule ringere end RF. Dette understreger begrænsningerne i modellering af *Culicoides* densitetsdata baseret på miljømæssige og klimatiske variable, og i denne afhandling blev der introduceret flere skævridninger, der følger af at samle data fra forskellige lande, der anvender forskellige indsamlingsprotokoller.

Resultaterne af denne afhandling tjener til at forstå den rumlige-temporale dynamik i hovedvektorerne for BT og SB i Europa, hvilket giver mulighed for forbedrede beslutninger med hensyn til sygdomsbekæmpelse og overvågningsprogrammer. At kende den rumlige vektorfordeling er et vigtigt skridt, der afgrænser mulige udbrud af vektorbårne sygdomme, da sygdommene ikke kan spredes ud over de områder, hvor vektoren er til stede. Derfor kan kortlægning af vektorfordelingen og tæthed bruges som et grundlæggende input til mere generelle modeller, der sammen med oplysninger om andre biologiske parametre kan hjælpe med at skabe R_0-kort, der er anvendelige i forbindelse med planlægning af både forebyggelse og kontrol af nye udbrud.

Resumen

En medicina veterinaria, las enfermedades transmitidas por *Culicoides* causan considerables pérdidas económicas, afectando el bienestar de los animales, reduciendo la producción y deteniendo el comercio de animales entre países donde la enfermedad esta presente y los que son libres de enfermedad. Otros costos indirectos están relacionados con los programas de vigilancia epidemiológica y programas de vacunación.

El virus de la lengua azul (BTV por sus siglas en inglés) y el virus de Schmallenberg (SBV) se transmiten a rumiantes (e.g. bovinos, caprinos y ovinos) a través de la picadura de jejenes hembras del género *Culicoides*. BTV se notificó en el sur de Europa durante el siglo XIX cuando apareció en algunos países de la cuenca mediterránea. En ese entonces, la ocurrencia de la enfermedad correspondía con la distribución de la especie afro-tropical *C. imicola*. BTV no se reportó nunca en el norte de Europa hasta el 2006, cuando comenzó un brote inesperado de la cepa BTV-8. La enfermedad se extendió y afectó varios países, causando grandes pérdidas económicas para la Unión Europea. Durante este brote, se iniciaron programas de vigilancia en muchos de los países afectados. Los resultados mostraron que *C. imicola* estaba completamente ausente en el área. A partir de esto se sospechó que las especies de los ensambles de Obsoletus y Pulicaris podrían transmitir el virus, lo cual se confirmó con la aislación de la cepa BTV - 8 a partir de jejenes silvestres capturados en campo. La enfermedad de Schmallenberg (SBV) fue descubierta en Alemania en 2011 y se extendió a 22 países europeos. Los brotes de esta enfermedad causaron malformaciones congénitas y muerte fetal en bovinos, ovinos y caprinos, lo que provocó pérdidas económicas.

Desde el comienzo de la epidemia de BTV, los estados miembros de la Unión Europea comenzaron a llevar a cabo programas de vigilancia entomológica para determinar las especies de vectores y la dinámica estacional. Varios países europeos han publicado los resultados de estos programas a nivel nacional. Algunos presentaron mapas de distribución de los vectores, a escala nacional. Sin embargo, existe aún la necesidad de generar mapas de abundancia y distribución del vector *Culicoides* a nivel continental, ya que la legislación europea sobre enfermedades transmitidas por vectores, se basa en decisiones conjuntas entre los

estados miembros. Por lo tanto, los datos entomológicos de los países europeos fueron recopilados por el proyecto VICE EMIDA (Infecciones transmitidas por vectores: sistemas de vigilancia rentables y basados en el riesgo) con el fin de crear una base de datos a nivel transnacional de *Culicoides*, para Europa. Nueve países acordaron compartir datos: España, Francia, Alemania, Austria, Suiza, Dinamarca, Suecia, Noruega y Polonia.

El objetivo de esta tesis fue analizar los datos disponibles, sobre *Culicoides*, recolectados en granjas europeas para comprender la dinámica espacial y temporal de los principales vectores de BTV y SBV en Europa, a escala continental. Recolectamos datos entomológicos existentes obtenidos en granjas de nueve países europeos, y se predijo la presencia y abundancia de estos vectores para Europa, desde el sur de España hasta el norte de Suecia. Todos los análisis se llevaron a cabo individualmente para los ensambles de Obsoletus y Pulicaris y para *C. imicola*.

En un primer paso, llevamos a cabo un análisis descriptivo de los datos entomológicos a nivel continental y exploramos la fluctuación temporal de la abundancia media considerando rangos latitudinales de 5° de ancho. También mapeamos la abundancia media mensual de las granjas muestreadas y usamos interpolación para detectar tendencias espaciales sobre la abundancia observada. Además, estimamos el inicio de la temporada a nivel NUTS para resaltar cualquier tendencia temporal. Los hallazgos principales de este análisis fueron que el ensamble de Obsoletus, el vector principal de BTV en la Europa Paleártica, mostró una tendencia sur-norte, dentro de la cual la mayor abundancia de vectores acumulados se encontró hacia el norte, aunque la duración de la temporada de vectores fue más corta en el norte. Esto podría ser el resultado de una adaptación del ensamble Obsoletus a áreas más frías.

En un segundo paso, modelamos los datos como presencia / ausencia utilizando la técnica de aprendizaje automático Random Forest (RF) basada en datos climáticos y datos ambientales de teledetección. Generamos mapas mensuales de la probabilidad de presencia de vectores. Estos mapas se clasificaron en tres clases según dos umbrales calculados para minimizar erroes en la clasificación. Las tres clases fueron: Presencia, Ausencia y una clase intermedia que representa las

áreas con un estado incierto, en donde se deberían enfocar esfuerzos de vigilancia. RF funcionó bien para el ensamble Obsoletus, razonable para el ensamble Pulicaris y muy bien para *C. imicola*. Los mapas presentados son útiles como herramienta básica para la toma de decisiones con respecto restricciones en el movimiento de animales o para la implementación de programas de vigilancia en esas áreas donde el modelo no fue capaz de clasificar las clases de presencia o ausencia.

Usando el mismo enfoque de modelado con datos climáticos y datos ambientales de teledetección, modelamos la abundancia por mes y creamos mapas mensuales de abundancia para los ensambles Obsoletus y Pulicaris y también para *C. imicola*. RF se desempeñó bien para predecir la abundancia del ensamble Obsoletus, razonable para el ensamble de Pulicaris y pobre para *C. imicola*, a pesar de que la distribución geográfica coincidió con la distribución conocida para esta especie. Cuando las predicciones son razonables, los valores predichos pueden ser útiles para la generación de modelos de evaluación de riesgos, por ejemplo, modelos R_0.

Por último, comparamos los resultados de RF con los de interpolación simple (algoritmo IDW) y los resultados mostraron que la interpolación fue ligeramente más pobre que el modelo RF. Esto resalta las limitaciones para modelar datos de abundancia de *Culicoides* basados en predictores ambientales y climáticos, y en el análisis de esta tesis, se introdujeron varios sesgos (bias) como consecuencia de la combinación de datos de diferentes países que utilizaron protocolos diferentes.

Los resultados presentados en esta tesis sirven para comprender la dinámica espacio-temporal del principal vector de BTV y SBV en Europa, lo que permite una mejor toma de decisiones para el control de enfermedades y creación de programas de vigilancia. Conocer la distribución espacial de vectores es importante porque delimita los límites de un posible brote de enfermedades transmitidas por vectores, y las enfermedades no pueden propagarse más allá de las áreas donde está presente el vector. Por lo tanto, el mapeo de la distribución y abundancia del vector puede usarse como un insumo fundamental para modelos más generales que, junto con la información de otros parámetros biológicos, podrían ayudar a crear mapas R_0 para las enfermedades transmitidas por *Culicoides*, útiles para mejorar la respuesta a posibles brotes de enfermedades.

Contents

Chapter 1

Introduction

1.1 Introduction

1.1.1 Vector borne disease and *Culicoides* biting midges

In an epidemiology context, a vector is a living organism that transmits a pathogen from one host to another. When we talk about "vector-borne diseases" we refer to a disease in which the pathogen, which could be a parasite, virus or bacteria, is transmitted to a host by a vector [OIE, 2011].

Vector-borne diseases affect not only humans (for example malaria and dengue), but also domestic animals. Therefore, the study of their transmission cycle is an important subject within the veterinary field. Outbreaks of vector-borne diseases have a negative economic impact on the livestock industry as they can cause a decrease in animal production, cause animal mortality or can lead to indirect financial losses such as restriction on animal trade, costs related to control measurements, surveillance monitoring programs or vaccination campaigns [EFSA, 2017].

Culicoides genus or "biting midges" are small (1-3 mm) blood sucking insects belonging to the order Diptera (Ceratopogonidae). There are more than 1300 species reported worldwide [Borkent, 2014] and the genus can be found occupying a wide range of habitats, with the exception of Antarctica, Patagonia and New Zealand [Mellor et al., 2000]. As any hematophagous insect, females seek for vertebrate hosts to get a blood meal in order to produce eggs. Their bites constitute a nuisance to humans, especially in certain areas of Europe (in UK). But

the importance of *Culicoides* lies in the fact that they are vectors of viruses that cause serious diseases in animals of economic importance.

In Europe, biting midges were responsible for the transmission and spread of animal diseases such as Bluetongue and Schmallenberg affecting cattle, goat and sheep, and African Horse sickness, affecting domestic horses and donkeys. In this thesis, we analysed the geographical distribution and abundance of *Culicoides* in Europe to be used as a basis for decision making to control diseases affecting livestock and therefore, the focus is on their veterinary importance.

1.1.2 The history of Bluetongue disease in Europe

Bluetongue disease (BT) is caused by an arbovirus (Reoviridae family), containing 27 recognized serotypes [Schulz et al., 2016]. The disease affects wild and domestic ruminants and among livestock, sheep presents more severe clinical signs compared to cattle and goats. The clinical symptoms in infected animals include fever, nasal discharge, facial and coronary bands hyperemia and edema, cyanotic lips or tongue, oral lesions and anorexia, lameness or pulmonary edema which can result in death.

From 1943, BT was reported occasionally in Cyprus, Spain, Portugal and Greece, but from 1998 the disease was finally introduced to the Mediterranean basin spreading towards France (Corsica), Italy, Turkey and countries of the Balkan region [Mellor et al., 2008, Sperlova and Zendulkova, 2011]. The situation worsened in August 2006 when a BTV-8 strain was reported in central Europe, (in the Netherlands) for the first time ever. Days later the disease was reported in Belgium, Germany, Luxemburg and northern France [Toussaint et al., 2007]. The following years BT spread even more towards southern and northern Europe affecting Denmark, Sweden and Norway, Switzerland, Austria, Hungary and Czech Republic [Ganter, 2014, Saegerman et al., 2008, Sperlova and Zendulkova, 2011, Wilson et al., 2007]. The BTV-8 epidemic is considered to be the most harmful of the BT epidemics of the last decades causing an estimated economic loss to the European Union of more than € 1000 million [Carpenter et al., 2009, 2013, Rushton and Lyons, 2015, Wilson and Mellor, 2008, Zientara and Sánchez-Vizcaíno, 2013].

Up to April 2018, 6 serotypes of BTV (1, 2, 3, 4, 6 and 16) were

circulating in Europe. The affected countries were Portugal, Spain, France, Switzerland, Italy, Malta, Slovenia, Croatia, Hungary, Romania, Bulgaria, Greece and Cyprus (Figure 1.1).

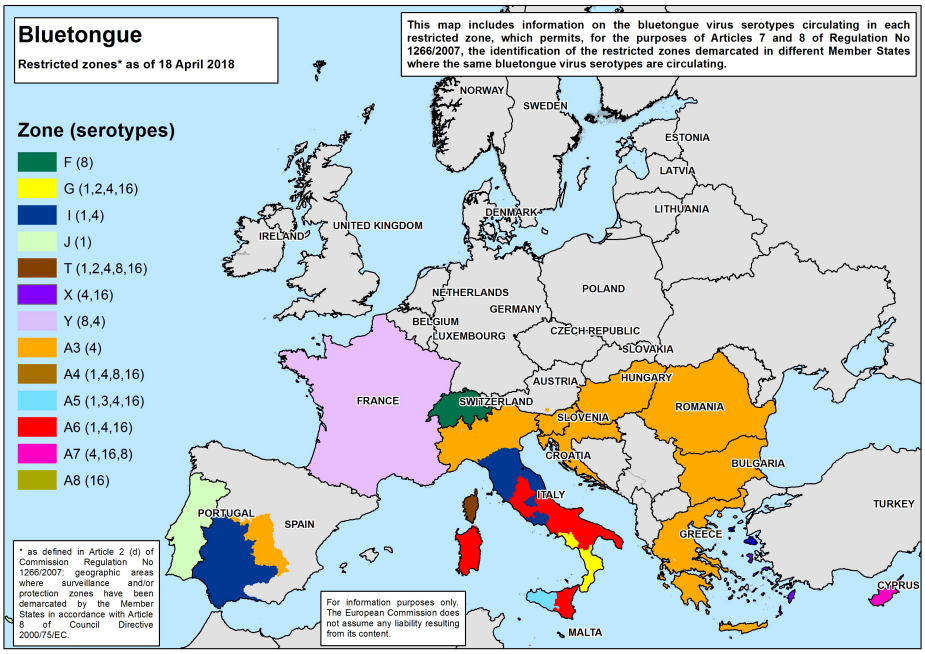


Figure 1.1: Countries affected by BTV on 18 April 2018 (Taken from https://ec.europa.eu/food/animals/animal-diseases/control-measures/bluetongue_en).

1.1.3 Schmallenberg

In August 2011, farmers from northwest Germany and the Netherlands reported the presence of cattle with clinical symptoms of a disease (fever, diarrhea, reduced milk yield). Sick animals were blood sampled and virus tests were carried out. A novel virus was then detected and named "Schmallenberg" virus (SBV) as the samples came from Schmallenberg city, in Germany [Afonso et al., 2014, Hoffmann et al., 2009]. SBV is a Orthobunyavirus (family Perunyaviridae) that, beside clinical symptoms in adult animals, may cause congenital malformations and still birth in cattle, sheep and goats. After the detection of the virus, Schmallenberg disease (SB) has since spread in Europe affecting 22 countries [Afonso

et al., 2014] (see Figure 1.2).

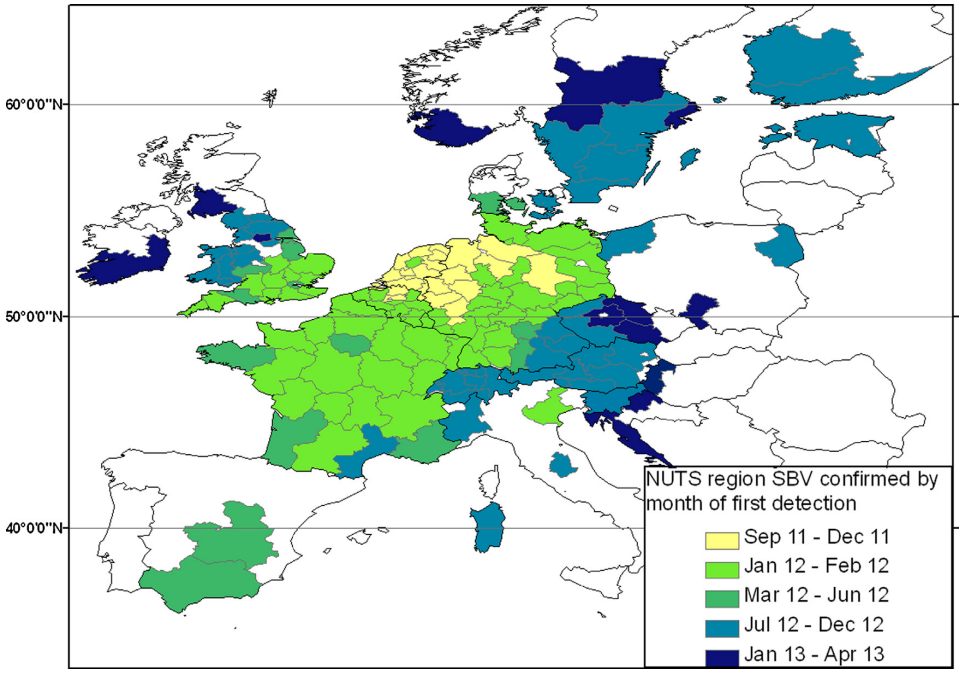


Figure 1.2: Regions (NUTS 2) with at least one SBV herd confirmed by direct detection by period of first report. Figure taken from Afonso et al. [2014].

1.1.4 The vectors of BT in Europe

Afro-tropical vector *C. imicola* was considered to be the vector of BT in southern European countries even though there was evidence of the role of endemic Palearctic species of *C. obsoletus* and *C. pulicaris* in BT transmission [Caracappa et al., 2003, Carpenter et al., 2009, Savini et al., 2005, Wilson and Mellor, 2008]. After 1998, *C. imicola* increased its geographical range from North Africa to the southern Mediterranean basin (Balearic Islands, Corsica, Sardinia, Sicily and southern Italy), possibly as a result of climate change [Purse et al., 2005]. But with the appearance of BT cases in central Europe, where *C. imicola* is completely absent, it became evident that the autochthonous *Culicoides* species of the Obsoletus complex were the ones transmitting the virus to ruminant livestock [Meiswinkel et al., 2008]. Entomological surveillance on European countries showed that specimens of *C. obsoletus* can be

found in high abundance while non *C. imicola* specimens were found in farms in NW Europe [EFSA, 2017, Meiswinkel et al., 2008, Sperlova and Zendulkova, 2011, Wilson and Mellor, 2008]. Other potential vectors of BT in Central and Northern Europe are *C. dewulfi* [Meiswinkel et al., 2008] and *Culicoides chiopterus* [Dijkstra et al., 2008, Venail et al., 2012].

1.1.5 Taxonomy of the European *Culicoides* vectors

The family Ceratopogonidae ("biting midges") comprises small hematophagous flies that, together with other families of hematophagous insects such as Culicidae (mosquitoes), Psychodidae (sand flies) are known vectors of pathogens to human and animals. The family is within the Diptera order (flying insects with only a functional pair of wings). In Europe, the species involved in the transmission of BTV and SBV belong to two subgenera: Avaritia and *Culicoides*. Avaritia comprises species of the Obsoletus group or complex: *Culicoides obsoletus*, *Culicoides scoticus*, *Culicoides montanus* and *Culicoides chiopterus* and the species *Culicoides dewulfi* and the afro-Asian *Culicoides imicola*. The subgenus *Culicoides* includes species of the so called Pulicaris group: *Culicoides pulicaris*, *Culicoides impunctatus* among others.

The current taxonomy of biting midges is in a "terrible condition" [Borkent, 2014]. According to this author, classification is mostly based on morphology leading to a wrong interpretation of phylogenetic relations among species. Such is the case of *C. dewulfi*, which is usually included within the Obsoletus group even though this species is not related to them. In this thesis, we opted to follow the proposal of Schwenkenbecher et al. [2009] of considering *C. dewulfi* as a species not included in the Obsoletus group based on molecular studies. Because the data obtained and used in this thesis consisted of abundance data of the Obsoletus group plus *C. dewulfi* gathered into a single group, we will use the term "ensemble" to refer to the group formed by the five species (equivalent to the Obsoletus group of other authors) and *C. dewulfi*. The word "ensemble" does not have a phylogenic meaning (Figure 1.3).

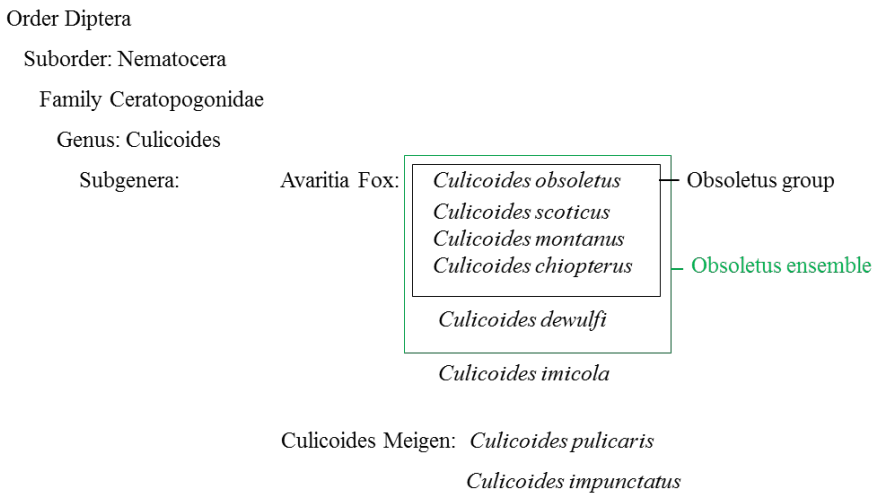


Figure 1.3: Scheme showing the taxonomy of the genus *Culicoides*. The black square remarks the species we consider included in the Obsoletus group and the green square remarks the species we consider as Obsoletus "ensemble" (with no phylogenetic meaning).

1.1.6 Species distribution modelling (SDM)

The geographic distribution of vectors determines the spatial boundaries of disease outbreak, as the vector is fundamental for transmission to occur. Identifying and mapping areas of vector occurrence is an important step for disease risk assessment. Spatial occurrence data (Presence / Absence) and / or abundance can be modelled using statistical methods to make predictions to non-sampled areas. These methods, called Species Distribution models (SDM), analyse the relationship between known occurrence records and predictor variables environmental factors) measured at the same localities of the vector occurrence, and create a model able to predict the response variable to un-sampled areas. The output of an SDM is a map or image showing the distribution of the species under study (Figure 1.4) showing the probability of presence (in case of modelling Presence / Absence data) or the abundance (if case of abundance data).

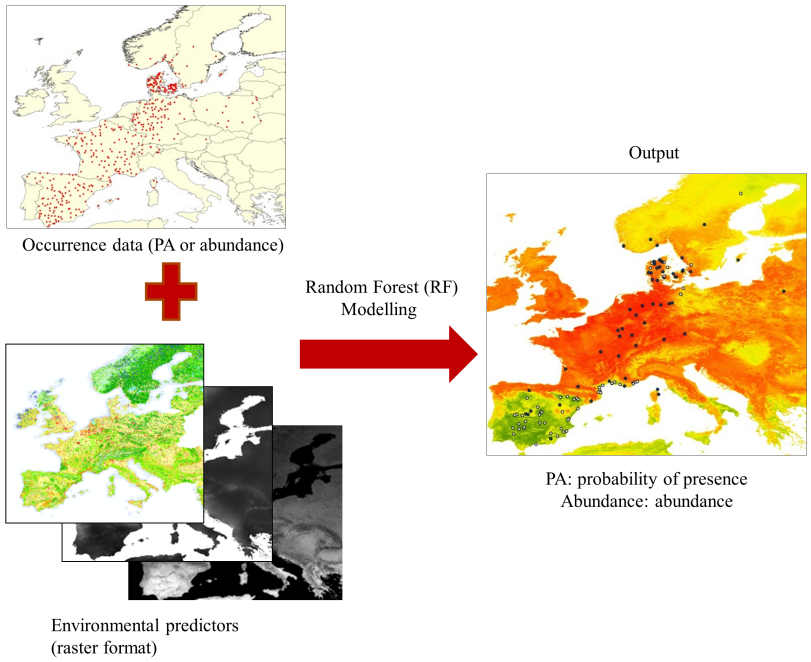


Figure 1.4: Scheme of a Species Distribution Modelling (SDM). SDM uses species data and environmental predictors to generate a distribution map, displaying probability of presence or abundance.

1.1.7 Vectors and their environment

The life cycle of *Culicoides* comprises four stages: egg, larva, pupa and adult. As they are cold blooded organisms, they are highly influenced by environmental conditions. Requirements for *Culicoides* development are dependent on the species but generally temperature and precipitation are the most important factors. An increase in temperature shortens the length of the stages of the life cycle and produces more generations over a period of time which makes adult populations grow faster [Mullens et al., 2004, Wittmann and Baylis, 2000]. Precipitation provides moisture for breeding sites, as most of the *Culicoides* larvae rear in very wet substrates [Kettle, 1962], even though *C. imicola* prefers dryer soils with organic matter, compared to other species of *Culicoides* [Mellor and Prrzous, 1979]. Additionally, precipitation is also related to humidity, a variable that affects the

level of activity and survival of adult specimens preventing them from desiccation [Purse et al., 2008]. Land cover in the surroundings has also been reported to affect the abundance of *Culicoides* for Obsoletus complex [Conte et al., 2007, De Liberato et al., 2010] and *C. imicola* [Tatem et al., 2003].

As environmental conditions and vector occurrence are related, environmental information can be used as predictors in statistical models. In the last decades, the use of remote sensing imagery as a surrogate of ground measurements has become more and more used in SMD studies. Some of the advantages of using remote sensing imagery are that they are easier to obtain (for instance, certain earth observation imagery can be download for free using the internet), they provide information of large regions at different pixel resolution and at different temporal frequency. If required, high resolution imagery can be obtained.

1.2 Outline of the thesis

In this thesis I work with the biggest data set aggregated for *Culicoides* in Europe to date. It includes entomological data collected in farms across nine European countries and comprises a transect of 4000 km from southern Spain to northern Sweden. This dataset, complemented with satellite imagery, containing information of environmental factors that affects *Culicoides* dynamics, constitutes the base materials for the understanding and elucidation of spatial and temporal patterns in the abundance, at a continental scale, of the main vectors of bluetongue and Schmallenberg diseases in Europe. The aims of this thesis were:

1. To understand the spatial pattern of *Culicoides* abundance on a monthly basis and at a continental scale.

2. To generate occurrence and abundance maps that can be used:

   (a) For decision making regarding the implementation of surveillance programs or regulations of animal movement between countries within the European Union.

(b) For risk assessments: as inputs in models that incorporates extra information of other parameters involved in the spread of vector borne diseases, such as temperature and virus occurrence. This will allow the generation of maps of risk of *Culicoides*-borne disease outbreaks (for instance R_0 maps).

To achieve this, the following objectives were developed:

1. To perform a descriptive analysis of the *Culicoides* abundance, analysing its temporal fluctuation by latitudinal ranges, and the start of the vector season by NUTS polygons.
2. To generate monthly maps of the vector occurrence in the nine countries and to divide the predicted probability of presence into classes useful for decision making regarding disease prevention and control.
3. To generate monthly maps of the vector abundance in the nine countries useful for future risk assessment based on R_0 models.

The results for each objective are found in each of the three manuscripts presented in chapter 3 (Results) of this thesis:

1. Manuscript I: "Spatial and temporal variation in the abundance of *Culicoides* biting midges (Diptera: Ceratopogonidae) in nine European countries. Published in Parasites & Vectors (2018). Using the observed abundance collected, we conducted a descriptive analysis in which we compared the seasonal fluctuation between different latitudes found in Europe. We also described, monthly and at a continent scale, the spatial variation of one *Culicoides* species (*C. imicola*) and two ensembles (Obsoletus and Pulicaris), highlighting the regions of high / low abundance and when the highest abundance occurred. Lastly, we analysed the start of the season at NUTS 3 level.
2. Manuscript II: "Predicting the average monthly *Culicoides* distribution and the vector-free seasons in nine European countries". Submitted to Parasites & Vectors. We model occurrence

data (the abundance is classified as Presence or Absence) using a machine learning technique (Random Forest (RF)) and predicted the probability of vector presence to non- sampled areas of Europe. We mapped the predictions obtained per each month, and these maps were classified into three classes (Absence, Uncertain and Presence). The classification maps can be used for decision making regarding the application of restriction of animal movements or the implementation of surveillance programs.

3. Manuscript III: "Modelling the monthly abundance of *Culicoides* biting midges in nine European countries using Random Forest machine learning techniques". In preparation for Parasites & Vectors. In manuscript III, we used the same data set and RF, but to predict the abundance of the vector to non-sampled areas in Europe. For each year sampled, an abundance map was obtained and we calculated the average of these maps to obtain a final average map per month. These maps are the first abundance maps created for Europe and they constitute key inputs for risk assessment modelling (such as R_0 models) [Gubbins et al., 2008, Hartemink et al., 2015] for BT, SB or emerging *Culicoides* diseases.

Chapter 2

Materials and Methods

This chapter describes how the entomological data was aggregated and the data management performed in order to obtain a compiled entomological data set. This single data set was then used to perform the different analyses in each manuscript. We also provide theoretical backgrounds for the methods employed for pre-processing the environmental predictors (Temporal Fourier Analysis) and for the modelling method used in Manuscripts II (section 3.2) and III (section 3.3) (Random Forest). The specific methodologies used in each analysis are described in section "Methods" within each manuscript (Chapter 3: Results).

2.1 *Culicoides* dataset

2.1.1 General description

European countries were contacted and asked to contribute with *Culicoides* catch data for this project. Nine countries agreed to share *Culicoides* data obtained from national surveillance programs and research projects: Spain [Acevedo et al., 2010], France [Balenghien et al., 2010, Venail et al., 2012], Germany [Mehlhorn et al., 2009], Switzerland [Kaufmann et al., 2012], Austria [Brugger and Rubel, 2013], Denmark, Sweden [Ander et al., 2012, Nielsen et al., 2010], Norway and Poland [Larska et al., 2013]. The data used here were published previously by each research group from each country with the exception of Denmark and Norway which have not published their results yet.

As the data did not come from a uniform and planned surveillance system in Europe, but was created by gathering entomological data available from different European projects, there were differences in the sampling protocols used by the countries, such as the number of trapping nights, the frequency of trapping in a month, the years sampled and the type of trap used.

Culicoides data consisted in catch data collected at cattle, sheep or/and horse farms using black light suction traps placed close to the stables. The data contained the following information: geographic coordinates of each farm recorded with GPS, start and end dates of trapping, number of nights sampled and type of trap used. For each observation, the specimens caught by a trap were identified to species level for *C. imicola* and to group level for specimens of the Obsoletus and Pulicaris group. As not all countries provided information of the gonotrophic stage of females specimens (nulliparous, parous, gravid), "number of *Culicoides*" is the total number of female *Culicoides* caught by a trap for each observation.

The sampling period varied among the countries and it ranged from 2007 to 2013. None of the countries were sampled less than two years or more than five. Traps were operated from dusk until sunset and, in most of the countries, were operational once a week. In Germany, traps were operational during seven nights in a row and emptied at the end of the sampling week. In Austria, only one site was sampled and the trap was operational and emptied every day for a three year period. Details of the sampling protocol followed by the individual countries are shown in Table 1 in Manuscript I (section 3.1).

2.1.2 Data management

Cleaning the data

The different countries sent their data as excel files which we compiled into a single large dataset. Data management included: Conversion of geographic coordinates to decimal degrees, transformation of coordinates to a single reference system (WGS84), correction of erroneous coordinates (e.g. mistakes in defining longitude and latitude), removal of records with missing coordinates, and the creation of unique IDs for each sample site.

Correcting for different trap types

All countries used Onderstepoort black light traps, except Germany and Spain, who used Biogents Sentinel (BG-Sentinel) traps (BioGents, Regensburg, Germany) and mini CDC model 1212 (John W. Hock, Gainesville, FL, USA), respectively. Light traps contain a light source (in this case black light) that attracts flying insects during the night. The trap is equipped with a fan that sucks the insects down to a collection beaker containing water that, when the observation is over, can be carried to the lab where specialists identify the species collected.

As Onderstepoort traps are reported to catch more specimens compared to BG-sentinel and mini CDC traps [Probst et al., 2015, Venter et al., 2009], we aimed to correct the trapping bias by using conversion factors. We multiplied the number of *Culicoides* caught in BG-Sentinel traps by 3.48 and for mini CDC traps, we multiplied the number of caught *Culicoides* by 2.5. Details on how these conversion factors were calculated can be found in the "Methods" section of Manuscript I (section 3.1).

2.2 Predictors used for *Culicoides* modelling

A raster is a grid in which each of the cells (called pixels) contains a numeric value representing information. Examples of raster formats are satellite imagery and all the products derived from them such as land cover maps and environmental indexes.

In this thesis, we used 112 raster files belonging to five different groups according to the information they provide and the source from where they were obtained. Collection and processing of the imagery is explained for each group:

1. **Fourier Transformed imagery**
   This raster data set consists of 70 satellite images derived from a temporal series of the MODIS sensor from 2001 to 2012. Five environmental variables were available: the infrared wavelength, night time land surface temperature, day time land surface temperature and two vegetation indexes: NDVI and EVI. These images were processed by the TALA group at Oxford University. The images were available to members of the EDENext project and were

downloaded using a FTP link.

Originally, for each of these variables, there were a set of MODIS satellite images taken at certain intervals of time from 2001-2012. Because the images in a temporal series are highly correlated [Rogers et al., 1996, Scharlemann et al., 2008], statistical utility of the imagery set is reduced. Temporal Fourier analysis (TFA) can be used to de-correlate the images and to reduce the long temporal series into a much shorter series of new raster images, keeping information about the seasonality. In satellites imagery, TFA is carried out pixel by pixel and all the values taken at different times within the temporal series are used to fit a series of sine functions that exhibit different curves or harmonics, different frequencies, amplitudes and phases and that collectively sum to the original time series. Each of these cycles/harmonics (the sine function) represents a seasonal cycle (annual, bi annual and tri annual) and their amplitude and phase have a biological interpretation. Amplitude component represents the variation of the cycle around the mean and the phase component, its timing (i.e. length of period cycles) [Rogers et al., 1996]. For each of the five environmental variables, 14 raster images are created from TFA, describing the overall mean (a0), amplitudes, phases and variance of each of the annual, bi-annual and tri-annual cycles (9 variables), the proportion of the variance of all three cycles combined (da); the maximum (mx) and minimum (mn) of the seasonal cycle recomposed from the first three harmonics only; and finally the variance (vr) of the original (i.e. not the fitted) time series [Hay et al., 2006].

Fourier transformed satellite imagery has been used previously to predict the distribution of tsetse flies in West Africa [Rogers et al., 1996], of *C. imicola* in Sicily [Purse et al., 2004] and the Mediterranean Basin [Baylis et al., 2002, Tatem et al., 2003, Wittmann et al., 2001], and species of the Obsoletus and Pulicaris groups in Scotland [Purse et al., 2012] and in the Mediterranean Basin [Purse et al., 2007]. It has also been used for predicting the occurrence of *Anopheles* mosquitoes in the Netherlands [Cianci et al., 2015, Ibañez-Justicia and Cianci, 2015].

2. **Worldclim dataset**
   Worldclim and Bioclim images are some of the most popular predictor variables used in SDM studies of terrestrial organisms. Worldclim is a dataset of global climate raster files with a spatial resolution ranging from 1 km^2 to 340 km^2. Worldclim provides temperature (min, mean and max) and precipitation data obtained from a temporal series from 1970 to 2000. Bioclim is dataset derived from Worldclim,that provides bioclimatic information from monthly temperature and precipitation. Bioclim rasters are available in the same spatial resolution as Worlclim images. "The bioclimatic variables represent annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, and precipitation of the wet and dry quarters). A quarter is a period of three months (1/4 of the year)." (Extracted from http://worldclim.org/bioclim). The 19 Bioclim images were downloaded from the online Worldclim database (http://worldclim.org/bioclim). The variables names are listed in Table 2 in Manuscript II (section 3.2).

3. **Corine Land Cover map**
   Corine Land Cover (CLC) is a map in raster format available in three spatial resolutions: 100 m^2, 250 m^2 and 1 km^2. CLC map provides information of 44 land cover classes in 12 European countries. CLC was downloaded from the European Environment Agency website. For this thesis, we used CLC map with a 250 m resolution. We selected 16 land cover classes and for each of them, we created binary maps (with values of 0 and 1) as dummy raster variables (Figure 2.1). From each class binary raster, we created new rasters calculating the frequency of the class within a bigger pixel (Figure 2.2). The details of this resampling process are described in the "Methods" section of Manuscript II (section 3.2). We present here two figures to complement the description of the methodology used on the CLC map in Manuscript II.

4. **Livestock density (FAO)**
   Raster files with livestock density data are available from the

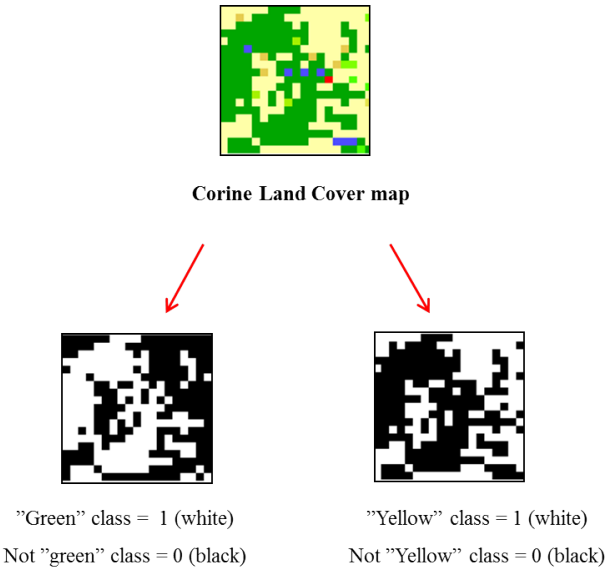


Figure 2.1: From an original CLC map, 16 classes were converted into binary raster files. This figure illustrates the procedure using only 2 classes as an example ("Green" class and "Yellow" class). For each binary file, the pixel where the class is present were given the value 1 (shown as white in the image) and pixels where the class is absent were given the value 0 (shown as black in the image)

Food and Agriculture Organization repository "GeoNetwork". We obtained 5 raster files of 1 km^2 resolution with information on cattle, goats, sheep, small ruminants and chicken.

5. **The "Altitude"** raster file (Digital Elevation Model) was obtained from Shuttle Radar Topography Mission (SRTM) (https://www2.jpl.nasa.gov/srtm/) of NASA.

For this thesis, 112 predictor variables were able to be used for fitting the RF models. One of the advantages of using machine learning (ML) is that they can handle large amounts of predictor variables,contrary to classical statistical methods such as Generalized Linear Models, in which variable selection is a previous step before modelling. Here, collinearity was assessed in the last manuscript only with the purpose of speeding the computation processing time.

Collinearity do not affect the modelling results. To demonstrate this, I compared two RF models, using the same training set for the Obsoletus ensemble for July (i) with the total 112 predictors and (ii) with only

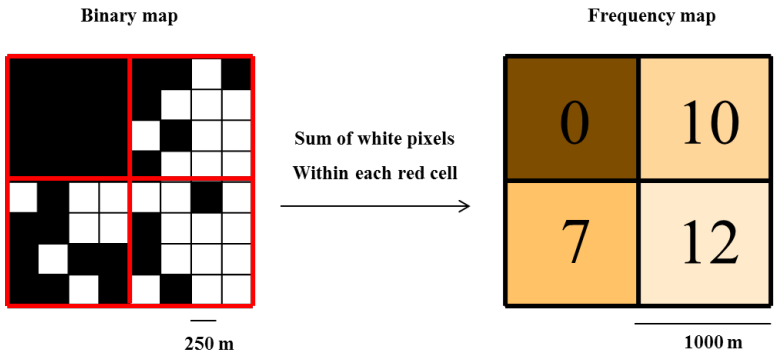


Figure 2.2: Each binary file were resampled and converted into a frequency map in the same step. In the figure an example is given using a subset (8 x 8 pixels) of a binary image. Each pixel of the binary image has a resolution of 250 m. A grid with the desired resolution (1 km) is then overlaid on the binary image (red grid). Each cell of the grid represents a pixel of the resulting resampled image and includes 16 pixels of the binary image. The new frequency map is created by summing the pixels with value 1 (white pixels) in each cell of the grid. The figure shows how from a subset of 8 x 8 pixels of 250 m resolution, we obtain a new image with 4 pixels (2 x 2) containing information of the frequency of the class for each 1km^2 cell.

85 predictors (removing from the analysis 25 predictors correlated to others).

To show this, I tested two different RF regression models using the mean abundance for the Obsoletus ensemble for July. One model was fit using the total of the predictors while the second models was fit using only 85 of the predictors. The test was predicted and I calculated the RMSE from the residuals. The results obtained showed a RMSE of 0,81 for the model without the set of predictors and a RMSE of 0,82 for the model including all the predictors. These results suggest that the amount of predictors do not affect the model performance.

2.3 Machine Learning techniques

Usually, data gathered by ecologists do not satisfy the assumptions required for being modelled with classic statistical methods, as the data show unusual distributions, dependence of the observations, non-linearity, zero-inflation (like the case of count data) and missing values [Crisci et al., 2012, Fielding, 1999]. These features pose a problem

for ecologists when trying to create predictive models, using classic statistics to classify or predict a response variable to non-sampled areas or periods as, if these assumptions are not satisfied, the models might lead to wrong results.

Machine Learning techniques (MLT) are a set of tools that can be used as surrogates of these classical statistical approaches for modelling ecological data (Fielding 1999) as they are able to handle data that do not satisfy the parametric model requirements. Ecological systems are mostly driven by complex interactions between possible predictor variables and classical statistical modelling might not able to find those complicated data patterns (e.g. generalized linear models) [Cutler et al., 2007]. Additionally, classical statistical models encounter problems with high numbers of predictor variables. Thus, MLT are suitable for ecological modelling as they are able to find complex and non-linear relations between the response variable and its predictors and they can handle a large amount of predictor variables. There are several ML algorithms and in this thesis, I used Random Forest (RF) (developed by Breiman [2001]) to predict the probability of presence and abundance of biting midges for every month. RF is based on an ensemble technique called Bagging and Classification and Regression Trees (CART). In this chapter, I give an introduction to the basic concepts for understanding how the Random Forest algorithm works and present the state of the art of this technique applied to *Culicoides* modelling in Europe.

2.3.1 Decision Trees

Decision trees is a method that recursively split data into two subsets (nodes) that are more homogenous than the parent node. At each node, the most important variable is chosen from all the predictors through a greedy algorithm and the best split point is computed [Hastie et al., 2009]. This procedure is repeated for each of the nodes of the tree until a stop criterion is reached (for example, certain amount of observations in the end node or when there is no change in the purity measurement). Once a tree is fully grown, there is a possibility that it may have over-fitted the data [Hastie et al., 2009, Kuhn and Johnson, 2013]. To avoid overfitting, the tree is pruned into a simpler tree using cost- complexity pruning (for further reading in the pruning algorithm see Hastie et al.

[2009], Kuhn and Johnson [2013]. The terminal nodes constitute the leaf nodes of the tree and each of them provide a prediction for new samples arriving to the terminal node. The nature of the predictions can be numerical, in the case of regression trees, or categorical for classification trees. The predictions are computed taking the mean value of all the observations belonging to a leaf (for regression) or assigning the class with the majority of votes (for classification) [Breiman, 2001, Cutler et al., 2007, Elith et al., 2008].

Decision trees have the advantage of being highly interpretable as trees are very easy to understand. Additionally, they can handle missing values and they are robust to outliers [Crisci et al., 2012]. Nevertheless, they present two major weaknesses: (1) they are unstable presenting higher variance. Unstable means that the structure of a tree is highly dependent on the samples constituting the training set. With an slightly change within the training data, the resulting trees might present a totally different structure [Elith et al., 2008, Kuhn and Johnson, 2013] and (2) their performance is suboptimal, as the relationship between the response variable and the predictor might not be defined by rectangular divisions of the prediction space. To overcome this issue, ensemble methods have been proposed. These methods use several models derived from the original data set and combine the predictions made by each model into a final prediction. An example of this method is Bagging.

2.3.2 Bagging

The term "bagging" refers to boostrap aggregation and it was proposed by Breiman in 1996. Bagging is an ensemble method that fits a decision tree to a bootstrap sample (sampling with replacement) taken from the original data set a repeated number of times (e.g. 500 or 1000). This result in a large amount of trees or "a forest". After the forest is created, a new sample being classified gets as many predictions as trees are present in forest. The final prediction for that new observation is computed as the average of all the predictions made by each tree, if regression trees are used, or from the majority of the votes in the case of classification trees [Breiman, 2001, Kuhn and Johnson, 2013]. The advantage of using bagging over single decision trees is that bagging reduces the variance of the predictions by using several subsamples of the original data set.

This makes the final predictions more stable, i.e. they are more robust to changes in the samples constituting the training data set [Kuhn and Johnson, 2013].

2.3.3 Random Forest

Trees created in bagging might be correlated (i.e. have similar structure) as all of them are generated analysing the total number predictor variables (M) when splitting each node of each tree. This affects bagging's power for reducing the variance. Leo Breiman presented the RF algorithm in 2001 [Breiman, 2001] which is the same as bagging but with an new incorporated feature: at each node of the tree, only on a random subset of the total number M of predictor variables is selected as possible candidate for splitting the node. This sophisticated feature decorrelates the trees in the forest by introducing an extra random factor to the tree construction. Thus, the tuning parameters in a RF model are the number of variables selected at each node (which is the same for all the nodes and all the trees) denoted as "m_{try}" and the total number of trees in the forest. Breiman recommends to set $m_{try} = M/3$ variables at each node for regression trees and $m_{try} = \sqrt{M}$ variables for classification trees, while Kuhn and Johnson, propose to tune this parameter using resampling techniques [Kuhn and Johnson, 2013].

Chapter 3

Results

3.1 Manuscript I

Spatial and temporal variation in the abundance of *Culicoides* biting midges (Diptera: Ceratopogonidae) in nine European countries

Ana Carolina Cuéllar, Lene Jung Kjær, Carsten Kirkeby, Henrik Skovgard, Søren Achim Nielsen, Anders Stockmarr,Mats Gunnar Andersson, Anders Lindström, Jan Chirico, Renke Lühken, Sonja Steinke, Ellen Kiel, Jörn Gethmann, Franz J. Conraths, Magdalena Larska, Inger Hamnes, Ståle Sviland, Petter Hopp, Katharina Brugger, Franz Rubel, Thomas Balenghien, Claire Garros, Ignace Rakotoarivony, Xavier Allène, Jonathan Lhoir, David Chavernac, Jean-Claude Delécolle, Bruno Mathieu, Delphine Delécolle, Marie-Laure Setier-Rio, Roger Venail, Bethsabée Scheid, Miguel Ángel Miranda Chueca, Carlos Barceló, Javier Lucientes, Rosa Estrada, Alexander Mathis, Wesley Tack and Rene Bødker.

Parasites & Vectors (2018) 11:112. doi: 10.1186/s13071-018-2706-y

Cuéllar et al. Parasites & Vectors (2018) 11:112
https://doi.org/10.1186/s13071-018-2706-y

Parasites & Vectors

RESEARCH Open Access

Spatial and temporal variation in the abundance of *Culicoides* biting midges (Diptera: Ceratopogonidae) in nine European countries

Ana Carolina Cuéllar[1*], Lene Jung Kjær[1], Carsten Kirkeby[1], Henrik Skovgard[2], Søren Achim Nielsen[3], Anders Stockmarr[4], Gunnar Andersson[5], Anders Lindstrom[5], Jan Chirico[5], Renke Lühken[6], Sonja Steinke[7], Ellen Kiel[7], Jörn Gethmann[8], Franz J. Conraths[8], Magdalena Larska[9], Inger Hamnes[10], Ståle Sviland[10], Petter Hopp[10], Katharina Brugger[11], Franz Rubel[11], Thomas Balenghien[12], Claire Garros[12], Ignace Rakotoarivony[12], Xavier Allène[12], Jonathan Lhoir[12], David Chavernac[12], Jean-Claude Delécolle[13], Bruno Mathieu[13], Delphine Delécolle[13], Marie-Laure Setier-Rio[14], Roger Venail[14,18], Bethsabée Scheid[14], Miguel Ángel Miranda Chueca[15], Carlos Barceló[15], Javier Lucientes[16], Rosa Estrada[16], Alexander Mathis[17], Wesley Tack[18] and Rene Bødker[1]

Abstract

Background: Biting midges of the genus *Culicoides* (Diptera: Ceratopogonidae) are vectors of bluetongue virus (BTV), African horse sickness virus and Schmallenberg virus (SBV). Outbreaks of both BTV and SBV have affected large parts of Europe. The spread of these diseases depends largely on vector distribution and abundance. The aim of this analysis was to identify and quantify major spatial patterns and temporal trends in the distribution and seasonal variation of observed *Culicoides* abundance in nine countries in Europe.

Methods: We gathered existing *Culicoides* data from Spain, France, Germany, Switzerland, Austria, Denmark, Sweden, Norway and Poland. In total, 31,429 *Culicoides* trap collections were available from 904 ruminant farms across these countries between 2007 and 2013.

Results: The Obsoletus ensemble was distributed widely in Europe and accounted for 83% of all 8,842,998 *Culicoides* specimens in the dataset, with the highest mean monthly abundance recorded in France, Germany and southern Norway. The Pulicaris ensemble accounted for only 12% of the specimens and had a relatively southerly and easterly spatial distribution compared to the Obsoletus ensemble. *Culicoides imicola* Kieffer was only found in Spain and the southernmost part of France. There was a clear spatial trend in the accumulated annual abundance from southern to northern Europe, with the Obsoletus ensemble steadily increasing from 4000 per year in southern Europe to 500,000 in Scandinavia. The Pulicaris ensemble showed a very different pattern, with an increase in the accumulated annual abundance from 1600 in Spain, peaking at 41,000 in northern Germany and then decreasing again toward northern latitudes. For the two species ensembles and *C. imicola*, the season began between January and April, with later start dates and increasingly shorter vector seasons at more northerly latitudes.
(Continued on next page)

* Correspondence: anacu@vet.dtu.dk
[1]Division for Diagnostics and Scientific Advice, National Veterinary Institute, Technical University of Denmark (DTU), Copenhagen, Denmark
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusion: We present the first maps of seasonal *Culicoides* abundance in large parts of Europe covering a gradient from southern Spain to northern Scandinavia. The identified temporal trends and spatial patterns are useful for planning the allocation of resources for international prevention and surveillance programmes in the European Union.

Keywords: *Culicoides* abundance, Seasonal abundance, Spatial pattern, Temporal trend, Vector season, *Culicoides* distribution, Europe, Vector-borne disease

Background

Biting midges of the genus *Culicoides* (Diptera: Ceratopogonidae) are important vectors of viruses among livestock, for example bluetongue virus (BTV), African horse sickness virus (AHSV) and Schmallenberg virus (SBV). The incursion of these viruses in Europe in recent decades has caused substantial economic losses to farmers in the European Union [1–8].

At least 83 species of *Culicoides* are found in Europe (83 species reported in France [9]) but only some of these are suspected to transmit viruses: the afrotropical vector *C. imicola* was previously considered to be the main vector of BTV in southern European countries [10], though BTV virus was also isolated from wild specimens of *Culicoides obsoletus* (Meigen)/*Culicoides scoticus* Downes & Kettle and specimens of the Pulicaris ensemble [11–14]. During an unprecedented outbreak of BTV serotype 8 in northern Europe in 2006, it became evident that autochthonous *Culicoides* species of the subgenus *Avaritia*, specifically *C. obsoletus/C. scoticus* and possibly *Culicoides dewulfi* Goetghebuer, were transmitting the virus [15–19].

In 2000, the European Commission established a series of regulations for BTV control, including monitoring and surveillance in the affected countries. According to Commission Regulation (EC) No. 1266/2007, it is mandatory for member states to carry out bluetongue surveillance programmes that include vector monitoring [20]. As a result of the BTV outbreak in 2006, the northern European countries also started carrying out entomological surveillance of *Culicoides* vectors. Collections by light traps at ruminant farms have been a key component of these programmes in both southern and northern Europe.

Several European countries have gathered and analysed entomological data at a national level to determine the presence and abundance of different species of *Culicoides* [21–25]. In addition, vector activity during the winter has been assessed in an attempt to determine the existence of a "vector-free period" [26, 27]. Determining a vector-free period might be useful for national veterinary authorities to authorize movements of test-negative ruminants. The number of national studies from European countries has increased during the last decade, but the need to quantify *Culicoides* vector dynamics at a continental level remains, as European legislation for vector-borne diseases is founded on joint decisions among the member states.

The aim of the present study was to generate a joint entomological database for large parts of Europe comprising different climatic zones, using surveillance and research data collected during the period 2007–2013. Nine European countries (Spain, France, Germany, Austria, Switzerland, Denmark, Sweden, Norway and Poland) agreed to share data and quantify key temporal and largescale geographical trends in the abundance of two main *Culicoides* species ensembles (Obsoletus ensemble and Pulicaris ensemble) and *C. imicola.*

Methods

Culicoides database

We gathered available *Culicoides* data from Spain [28], France [9, 23], Germany [24], Switzerland [29, 30], Austria [31], Denmark, Sweden [32, 33], Norway and Poland [34]. The data originated from national surveillance systems and research projects carried out during one or more years during a 7-year period (2007–2013) by national authorities and research groups. *Culicoides* biting midges were sampled from a total of 904 livestock farms (Fig. 1), with 31,429 trap collections comprising 8,842,998 specimens. For entomological sampling details from the different countries see [9, 23, 24, 28–34].

Black light suction traps were placed outside stables or near animal resting sites and the coordinates of each farm were recorded. Onderstepoort traps (Onderstepoort Veterinary Institute, Pretoria, Republic of South Africa) were used to catch *Culicoides* from dusk until dawn in all countries [23, 29, 31, 32, 34] except Germany, where Biogents Sentinel (BG-Sentinel) traps (BioGents, Regensburg, Germany) fitted with a black light lamp were used [24], and Spain, where mini CDC model 1212 (John W. Hock, Gainesville, FL, USA) traps were used [35]. Onderstepoort traps have been reported to catch more *Culicoides* than the other two types of traps [36, 37]. Therefore, data obtained by the BG-Sentinel and mini CDC traps were converted to the

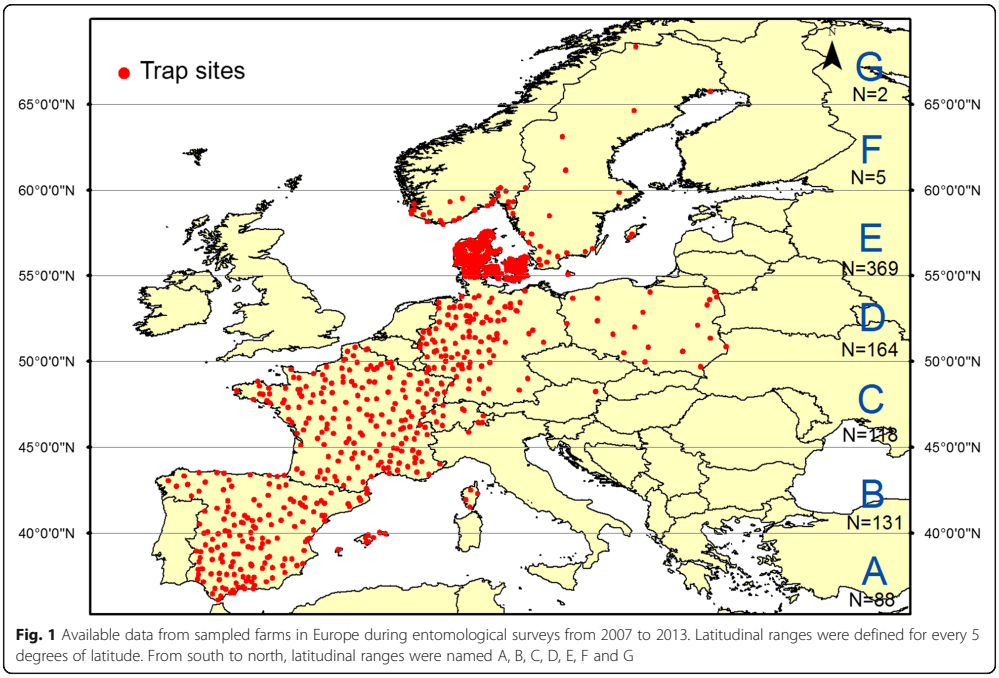


Fig. 1 Available data from sampled farms in Europe during entomological surveys from 2007 to 2013. Latitudinal ranges were defined for every 5 degrees of latitude. From south to north, latitudinal ranges were named A, B, C, D, E, F and G

number of specimens estimated to have been collected if Onderstepoort traps had been used. For BG-sentinel traps, Venter et al. [36] calculated a conversion factor of 3.1, while Probst et al. [37] calculated a conversion factor of 3.83. We used the mean of these values (3.48) as the conversion factor in this study. Three trap efficiencies (0.404, 0.288 and 0.505) were previously reported for the CDC mini trap [36] .We used the average of these values (0.399) and used the reciprocal conversion factor of 2.51 for all vector species. A single trap per farm was used in all the countries with the exception of Germany. During the 2012–2013 campaign Germany operated 3 traps per farm, so in order to have only one observation per farm we took the median amount of *Culicoides* caught among the 3 traps.

Some *Culicoides* species are difficult to identify based on morphology, e.g. *C. obsoletus*/*C. scoticus* [13, 18, 38–41]. In the data available for this analysis, specimens were identified to species level in some countries, while other countries only identified them to group level. To create a uniform database, we aggregated the species into ensembles. We use the term "ensemble" to refer to a group of sympatric species for which morphological identification is sometimes not possible or difficult, and without phylogenetic meaning. In this work, "Obsoletus ensemble" refers to the Obsoletus group and *C. dewulfi* together and includes the following species: *C. obsoletus*, *C. scoticus*, *Culicoides montanus* Shakirzjanova, *Culicoides chiopterus* (Meigen) and *C. dewulfi*. The Pulicaris ensemble includes *Culicoides pulicaris* (Linnaeus) and *Culicoides punctatus* (Meigen). As it is commonly used in the literature while authors differ in its composition, we refer to the Obsoletus group as a species group including *C. obsoletus*, *C. scoticus*, *C. montanus* and *C. chiopterus*. Based on the phylogenetic analysis of Schwenkenbecher et al. [42], we considered *C. dewulfi* as a separate species from Obsoletus group. We focused on these seven species, as they are considered to be farm-associated species [27, 29]. *Culicoides imicola* specimens were identified to species level by the two countries in which they were found (Spain and France). The sampling period is shown in Additional file 1: Table S1 while the number of trapping farms per country, trap type, and national protocols of the specific country are presented in Table 1.

European temperature data

We obtained daily temperature data from Europe between 1994 and 2004 from MARS-Agri4cast. As

Table 1 Number of farms sampled, number of collections, trap type used, frequency of trapping, sampling protocol per country and number of *Culicoides* specimens trapped (without applying conversion factor)

National survey	Details	No. of farms sampled	No. of collections	Trap type	Frequency	Sampling protocol (nights)	Total no. of Obsoletus ensemble	Total no. of Pulicaris ensemble	Total no. of *C. imicola*	Total no. of *Culicoides*
Austria		1	1095	Onderstepoort	Daily	1	16,338	1888	0	18,226
Denmark 1	2008–2009	343	1087	Onderstepoort	1 per month	1	193,795	174,475	0	368,270
Denmark 2	2010 winter surveillance	31	233	Onderstepoort	1–5 per month	1	297	222	0	519
Denmark 3	2012–2013 summer surveillance	4	102	Onderstepoort	1–5 per month	2 (2013); 3 (2014)	79,796	61,121	0	140,917
Total Denmark		350	1422	Onderstepoort	1–5 per month	1–3	273,888	235,818	0	509,706
France		192	10,947	Onderstepoort	1–5 per month	1	3,728,710	154,742	258,904	3,728,710
Germany 1	2007–2008	89	1244	BG-Sentinel	Monthly	7	901,235	203,101	0	1,104,336
Germany 2	3 campaigns: Aug-Sep 2012, April-May 2013, June 2013	21	664	BG-Sentinel	1 per period	14	46,895	38,465	0	85,360
Total Germany		110	1908	BG-Sentinel		7 and 14	948,130	241,566	0	1,189,696
Norway		29	698	Onderstepoort	1–10 per month	1	1,274,685	77,662	0	1,352,347
Poland		19	559	Onderstepoort	Weekly	1	277,546	160,751	0	438,297
Spain		168	12,724	Mini CDC	1–14 per month	1	254,331	35,799	196,324	486,454
Sweden		23	363	Onderstepoort	Weekly	1	60,052	14,979	0	75,031
Switzerland		12	1713	Onderstepoort	Weekly	1–2	489,789	141,091	0	630,880
Total		904	31,429				7,323,469	1,064,299	455,228	8,842,998

previously described by Beek [43], these data resulted from a linear interpolation of weather stations distributed across Europe into regular climate grids of 25 × 25 km.

Descriptive analysis and data management

We calculated the week number of each collection using the start date of the trapping. We defined week 1 of each year as 1st to 7th of January. This was done to ensure that the same dates from different years were given the same week number. We calculated the weekly and monthly mean abundance of vectors for each year. Finally, we calculated the average weekly and monthly abundance for the entire seven-year period to derive estimates for an "average year".

We conducted three different analyses. In the first analysis, we divided Europe into seven latitudinal bands (A-G) of 5° width, from 35°N to 70°N (Fig. 1). Latitude range G (> 65 °N) contained only two farms with just 9 observations from 4 weeks in August and September 2008, and was therefore not included in the latitudinal range analysis. To compare the seasonal variation among the seven different latitude ranges, we log transformed the trap collection data $[\log_{10}(x + 1)]$ and then calculated the mean of all the trap collections for each week number and at each latitude range, based on the data for the entire 7-year period (Fig. 1). To quantify the variation of the mean abundance for each week number, we calculated the 10th and 90th percentiles. For each latitudinal range, we also calculated the weekly average of the daily minimum, mean and maximum temperatures per week for the period from 1994 to 2004, and contrasted with the *Culicoides* seasonal variation. We estimated the number of vectors collected in an average year within each latitudinal range by calculating the annual cumulative sum of the weekly mean abundance and multiplying this by 7 days.

In the second analysis, we calculated the average abundance on each farm for each of the 12 months $[\log_{10}(x + 1)]$. We then spatially interpolated the log-transformed monthly averages to create spatial abundance maps. The interpolation was done using Inverse Squared Distance Weight (IDW) and based on the 15 nearest trap locations in ArcGIS 10.1 (ESRI, Redlands, CA, USA). We created buffer zones of 200 km around each farm and excluded all areas beyond this limit to avoid extrapolating to unsampled areas. In addition, countries outside the area of analysis were not included in the map.

In the third analysis we examined the spatial pattern of the start of the vector season by plotting the season start date of each NUTS (nomenclature of territorial units for statistics) 3rd level polygons defined by Eurostat (1992). NUTS is a hierarchical system used to divide up the economic territory of the EU for statistical purposes. We calculated the average start of the season for the 7-year period. The start of the season was defined as the first month in which the average monthly abundance per area polygon was equal to or higher than one specimen of *C. imicola,* and equal to or higher than 5 specimens of the Obsoletus and Pulicaris ensemble. The threshold numbers used here were based on (but not identical to) the threshold numbers defined by the European Commission to determine the start of the season [20]. While the EU thresholds are based on individual traps, we applied the thresholds to the average vector densities of all traps within a polygon. Polygons were classified as having "no data" if (i) the polygon did not have any sampled farms, (ii) if the mean abundance of the polygon did not reach the threshold during the year, or if (iii) there were no data available for the month prior the start of the season, thus making it impossible to detect whether the season might have started earlier.

Results

The trap data made available for our analyses included a total of 8,842,998 biting midges that had been collected at 904 farms between 2007 and 2013 in nine European countries. Of these, 82.8% belonged to the Obsoletus ensemble, 12.0% to the Pulicaris ensemble and 5.1% were *C. imicola*. Biting midges of the Obsoletus and Pulicaris ensemble were found in all countries that were sampled, while *C. imicola* was only found in Spain, along the southern coast of France and in Corsica.

Culicoides temporal fluctuation by latitude range

We observed a large variation in abundance in individual traps at each latitude range and at each week number, with the weekly 10th and 90th percentiles varying by a factor of 100 or more. However, when examining the average seasonal abundance, we were still able to observe two main patterns for the two species ensembles and *C. imicola*. First, the annual peak abundance in the Obsoletus ensemble increased gradually from Latitude A until it reached the highest weekly average peak of over 1000 vectors per night in Latitude F (Fig. 2, left column). Secondly, the period of the vector activity became increasingly shorter at higher latitudes, lasting throughout the whole year at Latitudes A and B (Fig. 2, left column), but only from week 15 (April) to week 46 (November) at Latitude F. Despite the increasingly shorter season for the Obsoletus ensemble further north, the cumulative sum of the weekly abundance steadily increased with latitude from less than 1000 Obsoletus ensemble vectors per year on average at Latitude A to 500,000 at Latitude F (Fig. 2, right column).

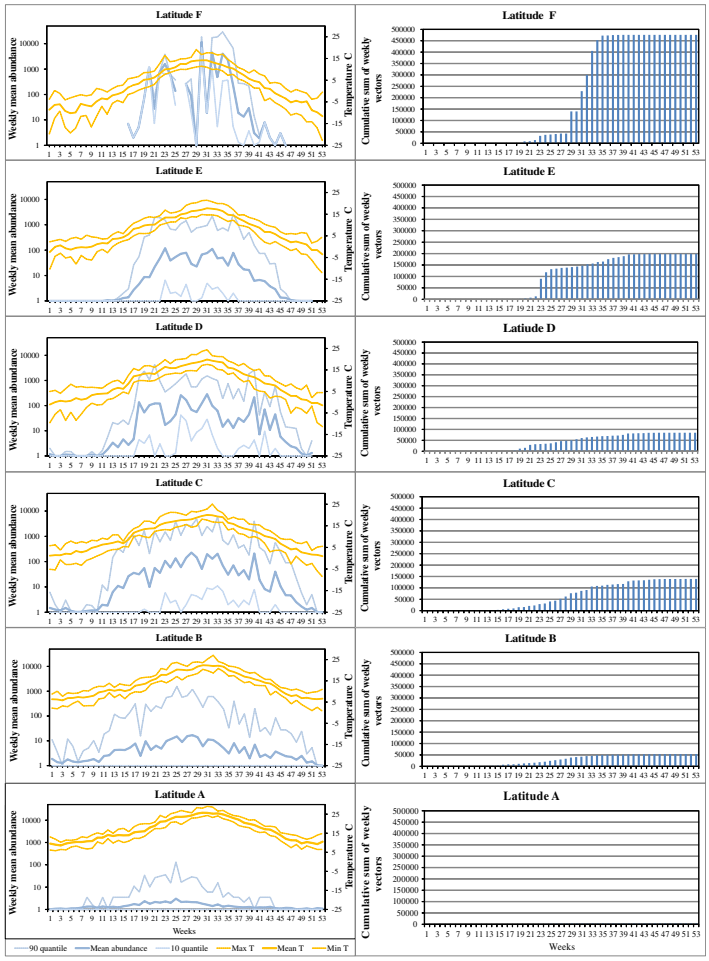


Fig. 2 Left column: Obsoletus ensemble weekly average (log scale) with 10th and 90th percentiles for an average year per latitudinal zone (A-F). Right column: cumulative weekly number of vectors per year, by latitudinal zone. The latitudinal zones ranged from southern Spain (A) to the northern Scandinavia (F)

We found a similar trend in the abundance of the Pulicaris ensemble vectors from south to north: the weekly mean abundance increased gradually from less than 5 in Latitude A to a peak weekly average of 100 in Latitude D (Fig. 3). However, in contrast to the Obsoletus ensemble, the Pulicaris ensemble abundance did not continue to increase beyond Latitude range D. In Latitude F, there was a marked variation in abundance across weeks (varying from low abundance one week to high peaks the next week). The cumulative sum of the weekly mean of Pulicaris ensemble abundance showed a different pattern compared to the Obsoletus ensemble, with the mean accumulated number of Pulicaris ensemble vectors peaking at mid-range latitudes (Latitude D) and reaching only 41,000 vectors per year (Fig. 3, right column). This value decreased to 15,000 vectors per year further north and south at Latitudes F and C, and to 1600 at Latitude A.

The duration of the season for the Pulicaris ensemble gradually decreased from Latitude B, where it lasted

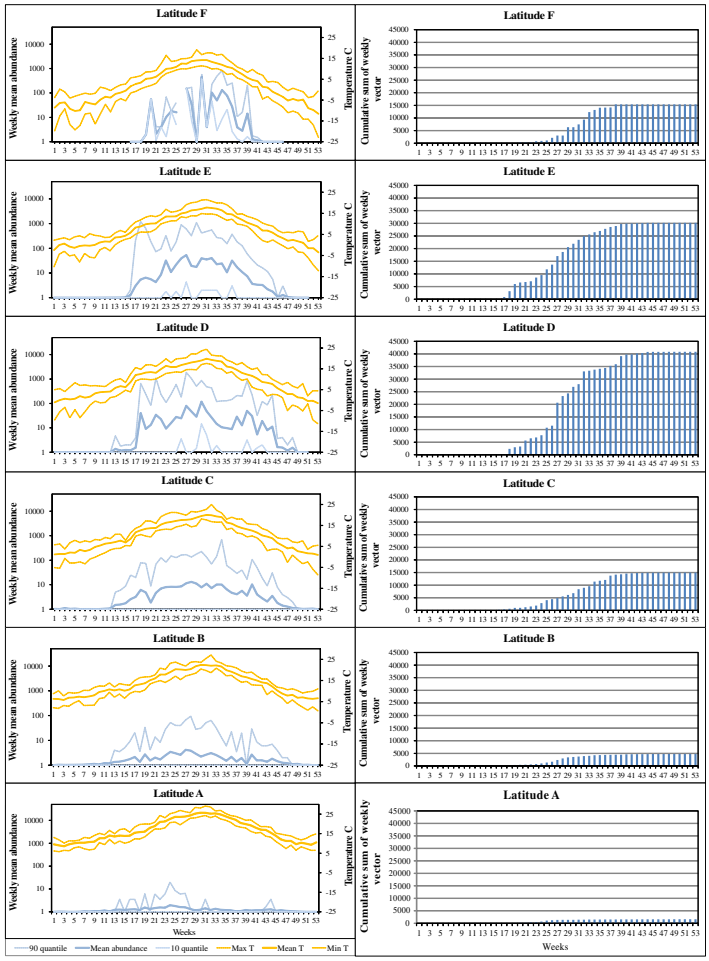


Fig. 3 Left column: Pulicaris ensemble weekly average (log scale) with 10th and 90th percentiles for an average year per latitudinal zone (A-F). Right column: cumulative weekly number of vectors per year, by latitudinal zone. The latitudinal zones ranged from southern Spain (A) to the northern Scandinavia (F)

from week 12 (April) to week 44 (October), toward Latitude F, where it lasted from week 18 (May) to week 41 (October). The vector season was shorter and the abundance lower for the Pulicaris ensemble than the Obsoletus ensemble at all latitude ranges. The mean temperature at the start of the vector season differed between Latitudes A and F. The season started with a mean temperature of 10 °C for the Obsoletus ensemble and 12 °C for the Pulicaris ensemble in southern latitudes, whereas the vector season started at much cooler mean temperatures (1 °C and 3 °C, respectively) at Latitude F (Figs. 2, 3; left columns).

At Latitude A, the abundance of *C. imicola* increased gradually until it reached the highest mean abundance of nearly 10 vectors per night (Fig. 4). At Latitude B (comprising northern Spain and Corsica), mean abundance was very low (< 3 specimens at the highest peak). This was due to *C. imicola* being almost absent in northern Spain at Latitude B, so the fluctuation of the observed abundance at this latitude was mainly caused by

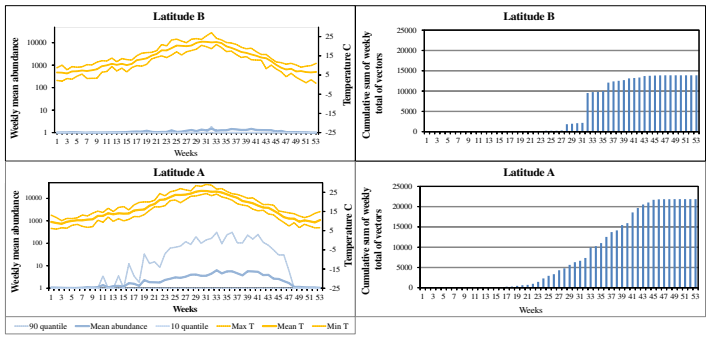


Fig. 4 Left column: *C. imicola* weekly average (log scale) with 10th and 90th percentiles for an average year per latitudinal zone (A-B). Right column: number of vectors per year, calculated as the cumulative sum of the weekly average multiplied by 7, by latitudinal range

collections made in Corsica. The vector season at Latitude A lasted from week 16 (April) to week 48 (November).

Abundance and interpolation maps

We observed a large variation in monthly abundance among farms in the same region for all two ensembles and *C. imicola*. However, spatially interpolation the mean monthly abundance at each farm revealed regions with a higher abundance and showed systematic variation within each latitude zone (Figs. 5 and 6). The regions with the highest monthly interpolated abundance of the Obsoletus ensemble were found in France (particularly in the north-west), Germany and southern Scandinavia (especially southern Norway). During the summer period, the monthly interpolated daily abundance was often high (1000–10,000) and occasionally very high (> 10,000) in these countries (Fig. 5). Every month, the interpolated abundance of the Pulicaris ensemble was of a lower magnitude than for the Obsoletus ensemble. During the summer months, farms with low (10–100) and medium (100–1000) abundance were often found distributed throughout the continent (with the exception of Spain), while the regions with the highest abundance were observed in Poland, Germany and Denmark with occasionally high-abundance (1000–10,000) farms (Fig. 6). The distribution in the high abundance regions were therefore found to be in more easterly areas for the Pulicaris ensemble compared to the Obsoletus ensemble.

The geographical areas where the Pulicaris ensemble and the Obsoletus ensemble showed the highest interpolated abundance during the summer months were generally also the areas where the species groups were observed earliest in the spring and latest in the autumn: western France for the Obsoletus ensemble, and Poland and Germany for the Pulicaris ensemble. In regions of low abundance, midges were first observed later and last observed earlier in the year (Figs. 5 and 6).

The highest abundance of *C. imicola* was found in Corsica, where a farm with extremely high abundance was found (> 10,000). In general, Spain had a medium abundance (100–1000), but high-abundance farms could occasionally be found. However, the abundance did not reach the levels seen for the Obsoletus ensemble in northern Europe (Fig. 7).

Start of vector season by NUTS 3 polygon level

We defined the start of the vector season for each NUTS 3 polygon as the first month with a mean abundance higher than or equal to one specimen for *C. imicola*, and higher than or equal to five specimens for the Obsoletus and Pulicaris ensembles.

According to this definition, the start of the season for the Obsoletus ensemble occurred as early as January in the west of France, some parts of Spain and Germany (Fig. 8) and as late as June in Scandinavia. In France, there was a clear spatial pattern where the Obsoletus ensemble season started early (January) in the west, and 2 to 3 months later in the east (March-April). In some provinces in Spain, the season started late (April-June).

The start of the vector season for the Pulicaris ensemble showed a spatial pattern similar to the Obsoletus ensemble, with a south-to-north gradient, where southern latitudes had an earlier start of the season (January to April in Spain and France) compared to northern latitudes (May to September in the Scandinavian countries). The start of the season for the Pulicaris ensemble occurred as early as January in some parts of Spain, Germany and Corsica, but generally occurred from March-April, 2 months later than for the Obsoletus ensemble (Fig. 9). In France, we observed the same pattern

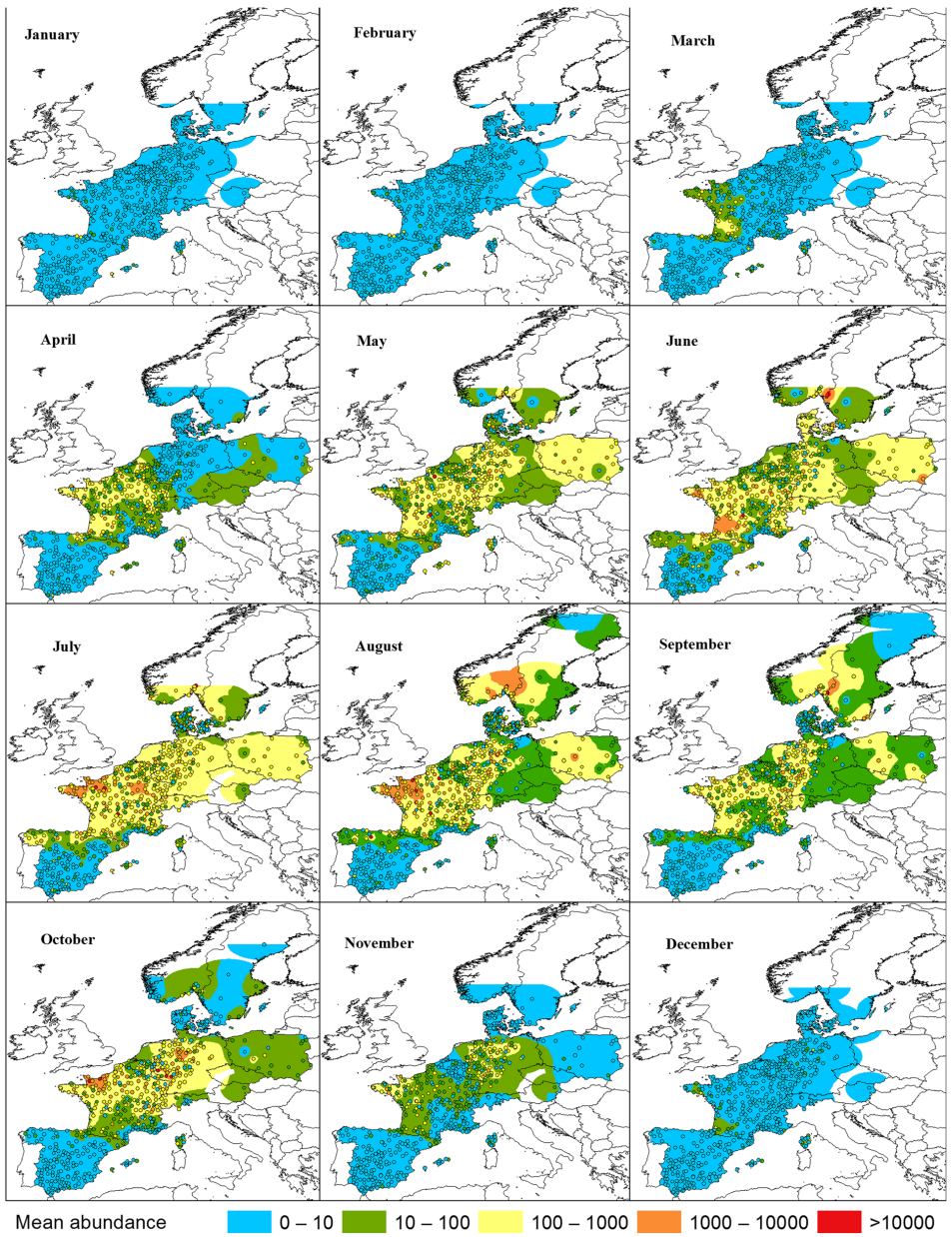


Fig. 5 Obsoletus ensemble monthly mean abundance. Dots show observed monthly mean abundance at sampled farms. Spatially interpolated abundance is shown in color. Interpolation values are displayed on the same scale as the observed abundance

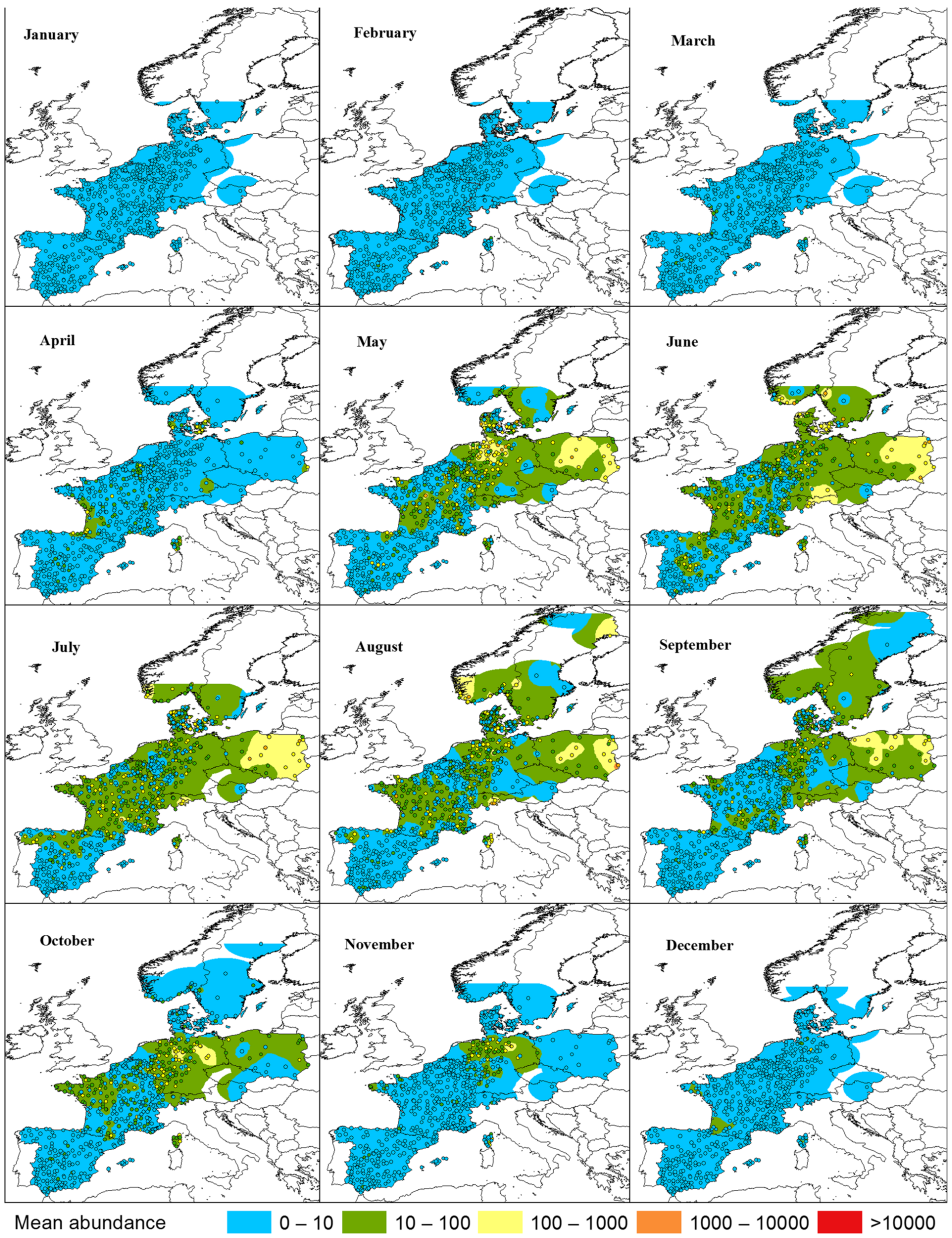


Fig. 6 Pulicaris ensemble monthly mean abundance. Dots indicate observed monthly mean abundance in sampled farms. Spatially interpolated abundance is shown in color. Interpolation values are displayed on the same scale as the observed abundance

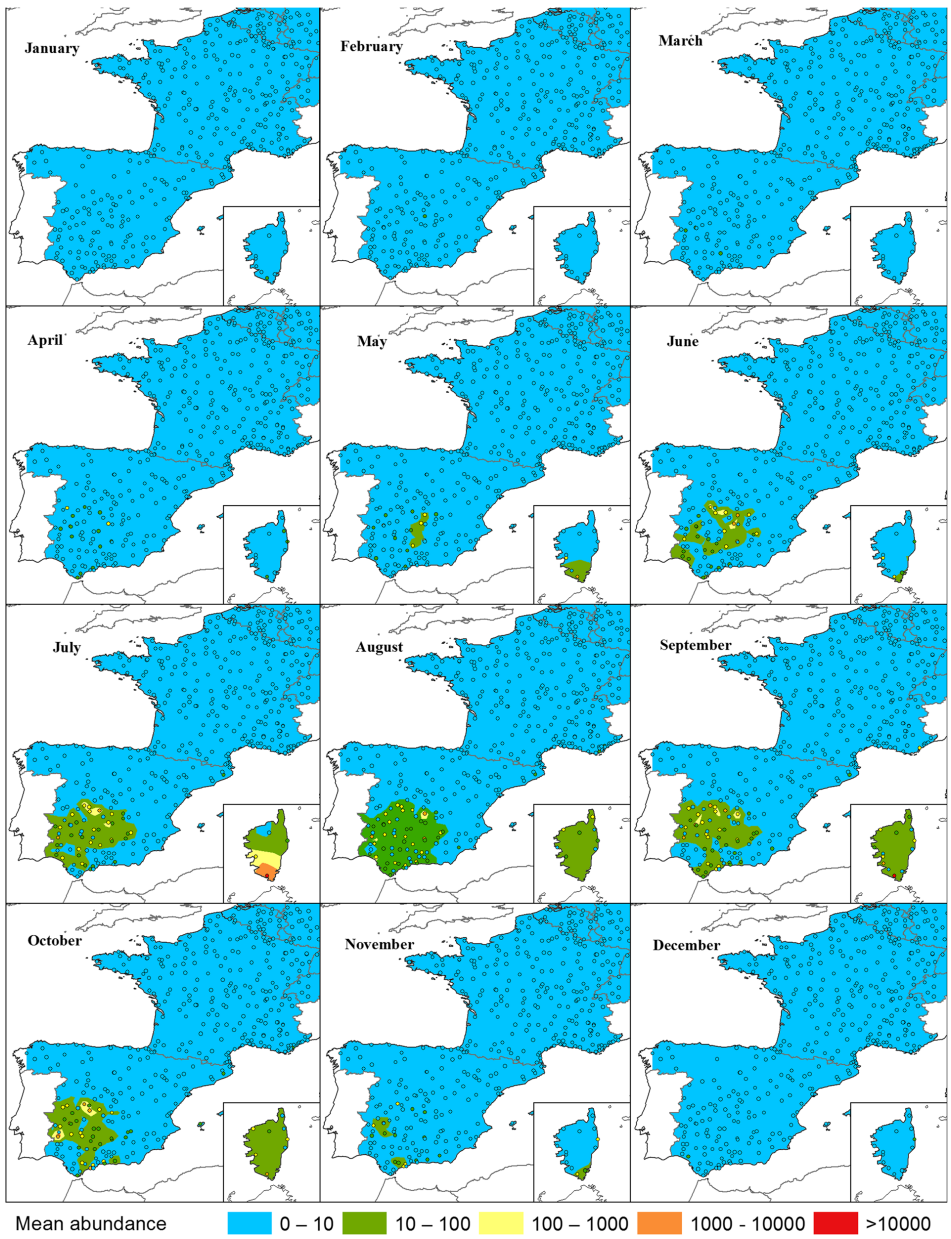


Fig. 7 *C. imicola* monthly mean abundance. Dots indicate observed monthly mean abundance in sampled farms. Spatially interpolated abundance is shown in color. Interpolation values are displayed on the same scale as the observed abundance

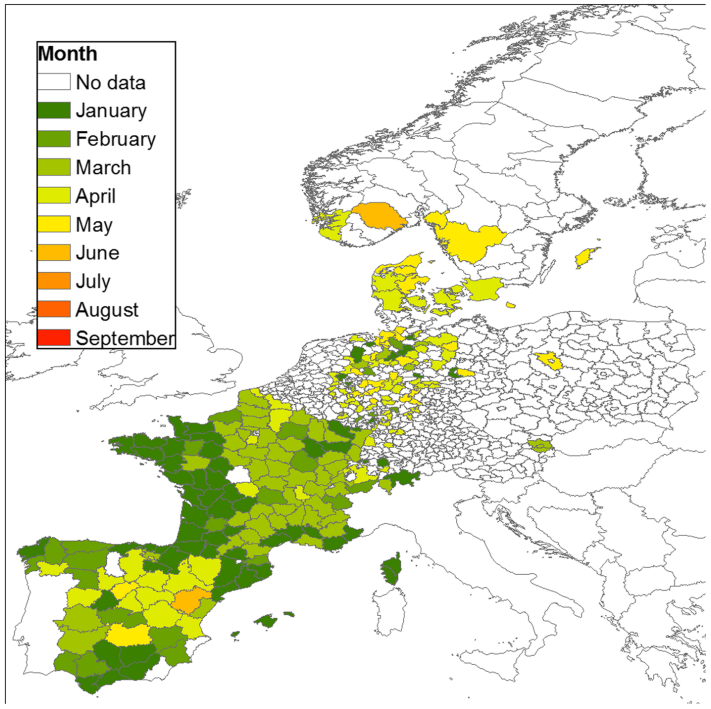


Fig. 8 Start of the vector season for the Obsoletus ensemble by NUTS 3 polygons and by month

found for the Obsoletus ensemble, where the start of the season for the Pulicaris ensemble occurred earlier (March) in the west compared to the eastern parts of the country (April).

Culicoides imicola was only recorded in Spain and France. The vector season started as early as January in southern Spain and on Corsica, while in the northern provinces of Spain, it started 5 months later (Fig. 10). However, there were also two provinces in the south of Spain where the average abundance per polygon did not reach the threshold value of one until June-August.

Discussion

The descriptive analysis presented here is based on the most extensive *Culicoides* dataset created for Europe to date, and represents the first combined description of *Culicoides* abundance and distribution for a large part of Europe. The data were gathered from a 4000 km long transect from southern Spain to the Arctic Circle in Sweden, with the most easterly collection sites in Poland. The primary aim of this descriptive analysis was to identify and quantify major geographical patterns and seasonal trends in the abundance of key *Culicoides* vector groups. The focus of the analysis was to identify patterns and trends important for decision making to prevent, surveillance and control of *Culicoides*-borne pathogens.

Specimens of the Obsoletus and Pulicaris ensemble were found in all of the countries sampled. The Obsoletus ensemble was ten times more abundant than the Pulicaris ensemble. Both groups have a Palaearctic distribution, and are widely distributed in Europe [13, 44–47]. However, the abundance of both ensembles and *C. imicola* varied dramatically among farms in the same region sampled during the same period, often showing a 100-fold difference between the 10th and the 90th percentile in terms of the weekly trap abundance within each latitude group. Nevertheless, distinct spatial and temporal patterns arose under the three analyses conducted in this work.

Examining the weekly data for the seven latitude ranges, we found that the mean weekly abundance of both ensembles and *C. imicola* varied dramatically along the south-north transect. It is interesting to note that the annual number of the Obsoletus ensemble gradually increased toward northern latitudes, despite the vector

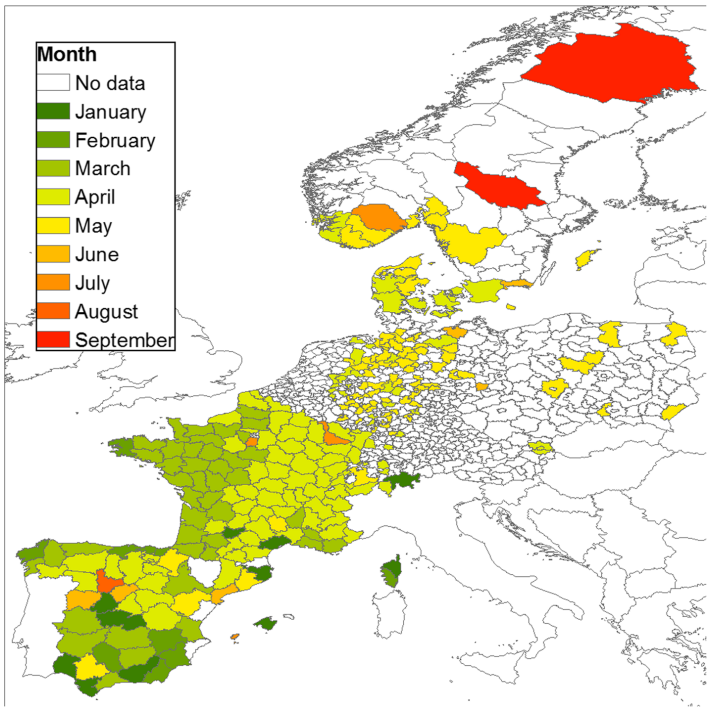


Fig. 9 Start of the vector season for the Pulicaris ensemble by NUTS 3 polygons and by month

season gradually shortening by three months from southern Spain to mid Scandinavia. This suggests that the Obsoletus ensemble is better adapted to the northern European climate in the environments surrounding farms. The Pulicaris ensemble appear better adapted to central Europe, and *C. imicola* to southern Europe, in the dry Mediterranean regions characterized by hot summers [47]. Although in relative terms, the Pulicaris ensemble (*C. pulicaris* and *C. punctatus*) was more abundant in central Europe and the Obsoletus ensemble (composed of *C. obsoletus*, *C. scoticus*, *C. montanus* and *C. chiopterus*.) was more abundant in northern Europe, the Obsoletus ensemble was still more abundant than the Pulicaris ensemble at all latitudes.

To further explore the spatial abundance patterns, we interpolated the mean monthly abundance for each calendar month. We found areas of high abundance peaks for the Obsoletus ensemble in north-western France, which is in agreement with previous studies that also reported high numbers of Obsoletus ensemble specimens in France [9, 22, 23]. The Obsoletus ensemble was also found in high abundance in most parts of Germany, where similar findings have been reported by many other authors [14, 19, 24, 47, 48]. Regarding the high-abundance areas found in Scandinavia in this study, these were partly driven by farms with an extremely high abundance; for example, two farms in southern Norway had more than 80,000 specimens per night (data not shown). Although the monthly mean data were $\log_{10}$ transformed before interpolation in order to reduce the impact of single sites of very high abundance, these high abundance farms from Norway still influenced the observed regional pattern. These high abundance records exceeded in great magnitude, the *C. obsoletus*/*C. scoticus* abundance previously reported in Sweden by Ander et al. [33] (> 5000 specimens in suitable months). Further collections in southern Norway are needed to determine how often the Obsoletus ensemble occurs at these extremely high numbers. The traps with the highest abundance of the Pulicaris ensemble were located in Germany and Poland, with a lower abundance in France and Scandinavia [14, 49, 50]. In general, the Pulicaris ensemble showed a spatial pattern with a relatively more easterly distribution compared to the Obsoletus

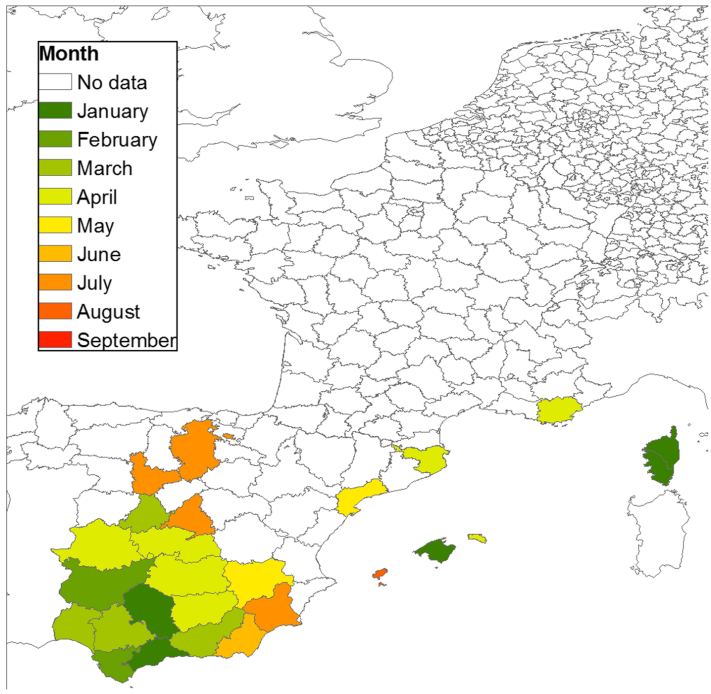


Fig. 10 Start of the vector season for *C. imicola* by NUTS 3 polygons and by months

ensemble. The distribution of the regions with the highest *C. imicola* abundance were in accordance with the known *C. imicola* distribution in Europe [9, 21, 22, 47]. Based on the interpolated monthly abundance maps, each vector group tended to appear early in the traps situated in areas that reached the highest abundance during summer, and were observed latest in the autumn. When the monthly maps are considered as a time series, the Obsoletus ensemble appears to start in western France, to increase in abundance and spread north and east until the end of August, when they retract to western France again. The Pulicaris ensemble appears to start in traps in Poland and, to a certain extent, the south-western areas of France, to grow in number and spread north and east before retracting again after August, to be observed last in Poland, northern Germany and western France. For *C. imicola*, the same phenomenon of areas with a strong spatial correlation between early appearance, peak abundance and a long vector season was seen in southern Corsica and south-eastern Spain.

The *Culicoides* abundance analysed in this study is exclusively based on the abundance observed on farms. The spatial abundance derived by interpolation therefore represents the abundance given the presence of a farm, and cannot be interpreted as an estimate of abundance in "non-farm" habitats, e.g. natural areas. Interpolation is used as tool to visualize the average abundance on farms in a larger area. To produce more detailed maps of vector abundance in future, a predictive modelling approach based on environmental and climate predictor variables will be necessary.

We analysed the start of the vector season for polygon areas in the participating countries. The European Commission currently defines the start of the vector season as the week during which the number of parous females exceeding a certain threshold (five for *C. obsoletus* and one for *C. imicola*) are caught by any trap in an area [20]. However, because there is a large variation in abundance among traps in the same area, the probability of finding a trap with an abundance exceeding the threshold will increase with the number of traps operated in the area. As a result, the vector season will tend to start earlier in regions with a higher number of traps. In this analysis, where the density of traps varies geographically, we used a more robust approach to define the start of

the vector season. We took into account the number of traps in a geographical area by calculating the mean trap abundance and using five vectors per trap as the cutoff (one female per trap for *C. imicola*). For the Obsoletus ensemble, there was not a clear pattern in Spain, where the vector season started at different months in different polygons. Nevertheless, there was a south-north trend starting in southern France and continuing to northern Scandinavia. The length of the vector season decreased by three months at northern latitudes where the period of climatic conditions suitable for midge development is shorter. Versteirt et al. [51] investigated the start and end of the vector season for *C. imicola* and *C. obsoletus*/*C. scoticus* on a continental scale in Europe. Their results also showed a south-north pattern in the start and length of the vector season. However, for *C. imicola*, our observed data suggested that the start of the season would occur earlier in the year (January–February) compared to their results where the season started in March. The start of the Pulicaris ensemble season had a similar pattern to the Obsoletus ensemble, starting two months later at higher latitudes; however, the start of the season seems to be more homogenous across latitudes compared to Obsoletus ensemble. We found the start of the season pattern for *C. imicola* is similar to previous results [51], with the only difference that in our study the vector season started as early as January on Corsica and in some provinces of Spain, which we interpret as areas with *C. imicola* vector activity all year round. The start of the season occurred later in the year in polygons where the peak *Culicoides* abundance in summer was low, as the density gradually increased and the cutoff defining the start of the vector season was reached later than in areas with a high peak abundance. For example, the Pulicaris ensemble activity started as late as September in northern Sweden, despite the first individual being observed much earlier. The start of the vector season is therefore highly sensitive to the vector abundance threshold selected to define it.

The latitude range analysis showed that the vector season for both the Obsoletus and Pulicaris ensembles started in the southern latitude range at an average mean temperature of 10 °C and 12 °C, respectively. However, at northern latitudes, the season started before spring temperatures rose, meaning that the vector season started at mean temperatures of just 1 °C and 3 °C. This suggests an adaptation of the vector population to the cooler climates in northern Europe. Early EU regulations suggested that vector-free periods may be defined by a specific temperature threshold when vector surveillance data were lacking [46]. Yet the start of the vector season at successively lower temperatures toward northern latitudes found in the present study demonstrates that such a simple temperature criterion alone would be a poor proxy for the start of the vector season, and would risk the prediction of an unrealistically long vector-free season in northern Europe.

One country used CDC traps (Spain) and one used BG-sentinel traps (Germany), while the rest used Onderstepoort traps. Previous studies compared the efficacy of different trap types, and the results showed that the number of midges collected varied according to the type used. It was reported that Onderstepoort traps collected more specimens than Mini CDC and BG-sentinel traps [35–37]. We attempted to adjust for this using published trap comparisons, but the relative efficiency of each type trap vary considerably. However, this uncertainty in trap efficiency is likely to be of a smaller magnitude than the variation in abundance within the spatial patterns which is of at least a 10 to 100 fold magnitude. The used trap conversion factor is therefore unlikely to have affected the identified overall spatial patterns identified here.

Light traps are most efficient when collection nights are dark [10, 52] and trapping in northern latitudes may therefore be less effective because nights during the summer season are shorter than in southern Europe [53]. The use of light traps in this study may therefore have underestimated the vector abundance at higher latitudes during the summer period.

We here focus on estimating the host seeking vector population. The vectors collected in light traps near stables and animal resting sites are likely to be predominantly host-seeking, while vectors that are already blood-fed are less likely to be collected in traps. The results from this analysis cannot directly be used for estimating the total vector population consisting of both host seeking and non-host seeking vectors. Because the blood meal digestion time in *Culicoides* is relatively slower at low temperatures [54], the proportion of the *Culicoides* population digesting blood and developing eggs and not attracted to traps may therefore be relatively larger at low temperatures. Due to the increasingly lower temperatures toward the north of the transect, the total vector population estimated from trap collections at higher latitudes, will be underestimated compared to the total vector population estimated from trap collections in southern Europe unless the blood meal developing time is taken into account.

Females of the species belonging to the Obsoletus group are difficult to separate based on morphological characters [13, 18, 38–41, 55] and therefore they are often grouped into the Obsoletus group or complex. The same occurs with the species of the Pulicaris group which are often merged into the Pulicaris group [25, 31, 56, 57]. Aggregating species into groups might represent a problem for identifying accurate temporal and spatial

patterns, as different species from the same group might exhibit different seasonal trends [58, 59]. The seasonal fluctuation of individual species of the Obsoletus group remains unknown. However, Searle et al. [27] analysed the phenology (start, end and duration of the vector season) of the male specimens for each species and the authors did not find a significant difference between the start and the end of season among the species. Nevertheless they found the length of the season period was different among them. Analyses regarding temporal fluctuation at species level of the Obsoletus group would be necessary as the species might present different temporal fluctuation patterns undetected when the species are grouped. Little is known about potential differences in vector competence for BTV for individual species of the Obsoletus and Pulicaris ensemble [11]. A study on the subject includes Carpenter et al. [60] who analysed vector competence for the Obsoletus group in the UK. Vector competence at species level of the Obsoletus group for SBV can be found in [60], Balenghien et al. [61] and Ségard et al. [62]. In this work, despite analyzing the data at ensemble level, we consider that the identified patterns and trends identified here still represent a useful and relevant overview of transmission potential in Europe.

Conclusions

This is the first report in which a dataset this size and covering a large part of Europe has been analysed. We identified and quantified the main mean spatial and temporal differences of three Culicoides species groups. Understanding the spatial and seasonal patterns of key vector groups or species facilitates the planning of preventive strategies and allows the development of more cost-effective vector and disease surveillance programmes by veterinary authorities in the European Union. The monthly abundance of the Obsoletus ensemble increased gradually from northern Spain to mid Scandinavia. The vector season also became increasingly shorter toward the north, starting three months later in mid Scandinavia compared to southern Spain. Nevertheless, the annual accumulated abundance of the Obsoletus ensemble increased steadily with latitude to 500,000 vectors per trap per year in mid Scandinavia. The Pulicaris ensemble was more frequent in central Europe, peaking in Germany and Poland with about 40,000 vectors per year, and with a more easterly distribution compared to the Obsoletus ensemble. For each of the species groups, there were areas in which the vectors appeared early, reached the highest mean peak abundances and lasted the longest. The Obsoletus ensemble was more abundant and had a longer season than the Pulicaris ensemble, whereas *C. imicola* appeared as a strictly southern species with a long vector season but with an abundance level that did not reach the peak abundance observed for the Obsoletus ensemble. This study suggests that future collaboration and data sharing between European countries may further improve our understanding of the spatio-temporal abundance of *Culicoides* vectors.

Additional file

Additional file 1: Table S1. Monthly availability of *Culicoides* trap data in the participating countries during the selected seven-year study period (2007–2013). X symbol indicates months when data were available. (XLSX 12 kb)

Abbreviations

BTV: bluetongue virus; NUTS: Nomenclature of territorial units for statistics; SBV: Schmallenberg virus

Acknowledgments

We would like to thank the Direction générale de l'alimentation from the French Ministry in charge of agriculture for funding and the Directions départementales de la protection des populations for their support in collecting the biting midges during the survey. We thank the Swiss Food Safety and Veterinary Office and the Vet-Austria project for financial support to the Swiss and Austrian partners, respectively. We also thanks the Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente for providing data about the national surveillance of *Culicoides* in Spain and the MARS-Agri4cast at the Institute for Environment and Sustainability, European Commission, for kindly providing temperature data of Europe.

Funding

This study was funded by the EMIDA ERA-NET-supported project VICE (Vector-borne Infections: risk-based and Cost-Effective surveillance systems). *Culicoides* data from Germany were partly collected within the German part of the VICE project funded by EMIDA ERA-NET through the Federal Office for Agriculture and Food (grant no. 314-06.01-2811ERA248).

Availability of data and materials

The national surveillance and research data that support the findings of this study are available from Spain: Miguel Ángel Miranda Chueca, France: Thomas Balenghien, Germany: Jörn Gethmann, Denmark: Rene Bødker, Sweden: Anders Lindström, Norway: Petter Hopp, Poland: Magdalena Larska, Austria: Katharina Brugger, Switzerland: Alexander Mathis but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of the national surveillance programmes of each country.

Authors' contributions

ACC analysed the data and drafted the manuscript. RB planned the original study, contributed to analysis and drafting the manuscript. LJK contributed to analysis and drafting the manuscript. CK, HS, SAN, AS, GA, AL, JC, RL, SS, EK, JG, FC, ML, IH, SS, PH, KB, FR, TB, CG, IR, XA, JL, JCD, BM, DD, MLS, RV, BS, MAMC, CB, JL, RE, AM and WT: discussed and identified preliminary national data, negotiated data access with national surveillance authorities and research projects, jointly discussed the taxonomic and spatial resolution for analysis and identified, and selected and extracted the final data and relevant variables meeting the criteria for the joint database, wrote the protocol summaries and commented on the analysis results and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Division for Diagnostics and Scientific Advice, National Veterinary Institute, Technical University of Denmark (DTU), Copenhagen, Denmark. [2]Department of Agroecology - Entomology and Plant Pathology, Aarhus University, Aarhus, Denmark. [3]Department of Science and Environment, Roskilde University, Roskilde, Denmark. [4]Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Copenhagen, Denmark. [5]National Veterinary Institute (SVA), Uppsala, Sweden. [6]Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Reference and Research National Reference Centre for Tropical Infectious Diseases, Hamburg, Germany. [7]Department of Biology and Environmental Sciences, Carl von Ossietzky University, Oldenburg, Germany. [8]Institute of Epidemiology, Friedrich Loeffler Institute, Greifswald, Germany. [9]Department of Virology, National Veterinary Research Institute, Pulawy, Poland. [10]Norwegian Veterinary Institute, Oslo, Norway. [11]Institute for Veterinary Public Health, Vetmeduni, Vienna, Austria. [12]CIRAD, UMR ASTRE, F-34398 Montpellier, France. [13]Institute of Parasitology and Tropical Pathology of Strasbourg, EA7292, Université de Strasbourg, Strasbourg, France. [14]EID Méditerranée, Montpellier, France. [15]Laboratory of Zoology, University of the Balearic Islands, Palma de Mallorca, Spain. [16]Department of Animal Pathology, University of Zaragoza, Zaragoza, Spain. [17]Institute of Parasitology, University of Zürich, Zürich, Switzerland. [18]Avia-GIS NV, Zoersel, Belgium.

Received: 5 December 2017 Accepted: 12 February 2018
Published online: 27 February 2018

References

1. Carpenter S, Wilson A, Mellor PS. *Culicoides* and the emergence of bluetongue virus in northern Europe. Trends Microbiol. 2009;17:172–8.
2. Carpenter S, Groschup MH, Garros C, Felippe-Bauer ML, Purse BV. *Culicoides* biting midges, arboviruses and public health in Europe. Antivir Res. 2013; 100:102–13.
3. Rushton J, Economic LN. Impact of bluetongue: a review of the effects on production. Vet Ital. 2015;51:401–6.
4. Wilson AJ, Mellor PS. Bluetongue in Europe: past, present and future. Philos Trans R Soc B. 2009;364:2669–81.
5. Zientara S, Sánchez-Vizcaíno JM. Control of bluetongue in Europe. Vet Microbiol. 2013;165:33–7.
6. Rasmussen LD, Kristensen B, Kirkeby C, Rasmussen TB, Belsham GJ, Bødker R, et al. Culicoids as vectors of Schmallenberg virus. Emerg Infect Dis. 2012;18: 1204–6.
7. Mellor PS, Boned J, Hamblin C, Graham S. Isolations of African horse sickness virus from vector insects made during the 1988 epizootic in Spain. Epidemiol Infect. 1990;105:447–54.
8. Pinior B, Brugger K, Kofer J, Schwermer H, Stockreiter S, Loitsch A, et al. Economic comparison of the monitoring programmes for bluetongue vectors in Austria and Switzerland. Vet Rec. 2015;176:464.
9. Venail R, Balenghien T, Guis H, Tran A, Setier-Rio M-L, Delécolle J-C, et al. Assessing diversity and abundance of vector populations at a national scale: example of *Culicoides* surveillance in France after bluetongue virus emergence. In: Mehlhorn H, editor. Arthropods as vectors of Emerging Diseases. Heidelberg: Springer-Verlag; 2012. p. 77–102.
10. Carpenter S, Szmaragd C, Barber J, Labuschagne K, Gubbins S, Mellor P. An assessment of *Culicoides surveillance* techniques in northern Europe: have we underestimated a potential bluetongue virus vector? J Appl Ecol. 2008; 45:1237–45.
11. Savini G, Goffredo M, Monaco F, Di Gennaro A, Cafiero MA, Baldi L, et al. Bluetongue virus isolations from midges belonging to the Obsoletus complex (*Culicoides*, Diptera: Ceratopogonidae) in Italy. Vet Rec. 2005;157: 133–9.
12. De Liberato C, Scavia G, Lorenzetti R, Scaramozzino P, Amaddeo D, Cardeti G, et al. Identification of *Culicoides obsoletus* (Diptera: Ceratopogonidae) as a vector of bluetongue virus in central Italy. Vet Rec. 2005;156:301–4.
13. Caracappa S, Torina A, Guercio A, Vitale F, Calabrò A, Purpari G, et al. Identification of a novel bluetongue virus vector species of *Culicoides* in Sicily. Vet Rec. 2003;153:71–4.
14. Clausen PH, Stephan A, Bartsch S, Jandowsky A, Hoffmann-Köhler P, Schein E, et al. Seasonal dynamics of biting midges (Diptera: Ceratopogonidae, *Culicoides* spp.) on dairy farms of central Germany during the 2007/2008 epidemic of bluetongue. Parasitol Res. 2009;105:381–6.
15. Wilson a CS, Gloster J, Mellor P. Re-emergence of bluetongue in northern Europe in 2007. Vet Rec. 2007;161:487–9.
16. Sperlova A, Zendulkova D. Bluetongue: a review. Vet Med (Praha). 2011;56:430–52.
17. Meiswinkel R, Baldet T, de Deken R, Takken W, Delécolle J-C, Mellor PS. The 2006 outbreak of bluetongue in northern Europe - the entomological perspective. Prev Ve Med. 2008;87:55–63.
18. Nolan DV, Carpenter S, Barber J, Mellor PS, Dallas JF, Mordue Luntz AJ, et al. Rapid diagnostic PCR assays for members of the *Culicoides obsoletus* and *Culicoides pulicaris* species complexes, implicated vectors of bluetongue virus in Europe. Vet Microbiol. 2007;124:82–94.
19. Vorsprach B, Meiser CK, Werner D, Balczun C, Schaub GA. Monitoring of Ceratopogonidae in southwest Germany. Parasitol Res. 2009;105:337–44.
20. European Commission. EC 1266/2007. Off J Eur Union. 2007;L 283 of 27.10. 2007:37–52.
21. Calvete C, Miranda MA, Estrada R, Borras D, Sarto i, Monteys V, Collantes F, et al. Spatial distribution of *Culicoides imicola*, the main vector of bluetongue virus, in Spain. Vet Rec. 2006;158:130–1.
22. Balenghien T, Garros C, Mathieu B, Allène X, Gardes L, Rakotoarivoany I, et al. La surveillance des *Culicoides* en France. Bull Epidémiol Santé Anim Aliment. 2010;35:8–9.
23. Balenghien T, Delécolle J, Rakotaoarivony I. Bluetongue - report on entomological surveillance in France in 2010. Bull Epidémiol Santé Anim Aliment. 2010;46:26–31.
24. Mehlhorn H, Walldorf V, Klimpel S, Schaub G, Kiel E, Focke R, et al. Bluetongue disease in Germany (2007–2008): monitoring of entomological aspects. Parasitol Res. 2009;105:313–9.
25. Goffredo M, Conte A, Distribution MR. Abundance of *Culicoides imicola*, Obsoletus complex and Pulicaris complex (Diptera: Ceratopogonidae) in Italy. Vet Ital. 2004;40:270–3.
26. Brugger K, Köfer J, Rubel F. Outdoor and indoor monitoring of livestock-associated *Culicoides* spp. to assess vector-free periods and disease risks. BMC Vet Res. 2016;12:88.
27. Searle KR, Barber J, Stubbins F, Labuschagne K, Carpenter S, Butler A, et al. Environmental drivers of *Culicoides* phenology: how important is species-specific variation when determining disease policy? PLoS One. 2014;9: e111876.
28. Acevedo P, Ruiz-Fons F, Estrada R, Márquez AL, Miranda MA, Gortázar C, et al. A broad assessment of factors determining *Culicoides imicola* abundance: modelling the present and forecasting its future in climate change scenarios. PLoS One. 2010;5:e14236.
29. Kaufmann C, Steinmann IC, Hegglin D, Schaffner F, Mathis A. Spatio-temporal occurrence of *Culicoides* biting midges in the climatic regions of Switzerland, along with large scale species identification by MALDI-TOF mass spectrometry. Parasit Vectors. 2012;5:246.
30. Kaufmann C, Schaffner F, Mathis A. Monitoring von Gnitzen (*Culicoides* spp.), den potentiellen Vektoren des Blauzungenkrankheitsvirus, in den 12 Klimaregionen der Schweiz. Schweiz Arch Tierheilkd. 2009;151:205–13.
31. Brugger K, Rubel F. Bluetongue disease risk assessment based on observed and projected *Culicoides obsoletus* spp. vector densities. PLoS One 2013;8: e60330.
32. Nielsen SA, Nielsen BO, Chirico J. Monitoring of biting midges (Diptera: Ceratopogonidae: *Culicoides* Latreille) on farms in Sweden during the emergence of the 2008 epidemic of bluetongue. Parasitol Res. 2010;106: 1197–203.
33. Ander M, Meiswinkel R, Chirico J. Seasonal dynamics of biting midges (Diptera: Ceratopogonidae: *Culicoides*), the potential vectors of bluetongue virus, in Sweden. Vet Parasitol. 2012;184:59–67.
34. Larska M, Lechowski L, Grochowska M, Żmudziński JF. Detection of the Schmallenberg virus in nulliparous *Culicoides obsoletus/scoticus* complex and *C. punctatus* - the possibility of transovarial virus transmission in the midge population and of a new vector. Vet Microbiol. 2013;166:467–73.
35. del Río R, Monerris M, Miquel M, Borràs D, Calvete C, Estrada R, et al. Collection of *Culicoides* spp. with four light trap models during different seasons in the Balearic Islands. Vet Parasitol. 2013;195:150–6.

36. Venter GJ, Labuschagne K, Hermanides KG, Boikanyo SNB, Majatladi DM, Morey L. Comparison of the efficiency of five suction light traps under field conditions in South Africa for the collection of *Culicoides* species. Vet Parasitol. 2009;166:299–307.
37. Probst C, Gethmann JM, Kampen H, Werner D, Conraths FJ. A comparison of four light traps for collecting *Culicoides* biting midges. Parasitol Res. 2015; 114:4717–24.
38. Goffredo M, Entomological MR. Surveillance of bluetongue in Italy: methods of capture, catch analysis and identification of *Culicoides* biting midges. Vet Ital. 2004;40:260–5.
39. Baldet T, Delécolle JC, Cêtre-Sossah C, Mathieu B, Meiswinkel R, Gerbier G. Indoor activity of *Culicoides* associated with livestock in the bluetongue virus (BTV) affected region of northern France during autumn 2006. Prev Vet Med. 2008;87:84–97.
40. Mathieu B, Perrin A, Baldet T, Delécolle J-C, Albina E, Cêtre-Sossah C. Molecular identification of western European species of obsoletus complex (Diptera: Ceratopogonidae) by an internal transcribed spacer-1 rDNA multiplex polymerase chain reaction assay. J Med Entomol. 2007;44:1019–25.
41. Kluiters G, Pagès N, Carpenter S, Gardès L, Guis H, Baylis M, et al. Morphometric discrimination of two sympatric sibling species in the Palaearctic region, *Culicoides obsoletus* Meigen and *C. scoticus* Downes & Kettle (Diptera: Ceratopogonidae), vectors of bluetongue and Schmallenberg viruses. Parasit Vectors. 2016;9:262.
42. Schwenkenbecher JM, Mordue AJ, Piertney SB. Phylogenetic analysis indicates that *Culicoides dewulfi* should not be considered part of the *Culicoides* obsoletus complex. Bull Entomol Res. 2009;99:371–5.
43. Beek EG. Spatial interpolation of daily meteorological data. Theoretical evaluation of available techniques. 1991. http://edepot.wur.nl/360405. Accessed 17 Feb 2018.
44. Hill MA. The life-cycle and habits of *Culicoides impunctatus* Goetghebuer and *Culicoides obsoletus* Meigen, together with some observations on the life-cycle of *Culicoides odibilis* Austen, *Culicoides pallidicornis* Kieffer, *Culicoides cubitalis* Edwards and *Culicoides chiopterus* Meigen. Ann Trop Med Parasitol. 1947;41:55–115.
45. Purse B, Brown HE, Harrup L, Mertens PPC, Invasion RDJ. Of bluetongue and other orbivirus infections into Europe: the role of biological and climatic processes. Rev Sci Tech. 2008;27:427–42.
46. EFSA Panel on Animal Health and Welfare. Bluetongue: control, surveillance and safe movement of animals. EFSA J. 2017;15
47. Brugger K, Rubel F. Characterizing the species composition of European *Culicoides* vectors by means of the Köppen-Geiger climate classification. Parasit Vectors. 2013;6:333.
48. Hörbrand T, Geier M. Monitoring of *Culicoides* at nine locations in southern Germany (2007–2008). Parasitol Res. 2009;105:387–92.
49. Kiel E, Liebisch G, Focke R, Liebisch A. Monitoring of *Culicoides* at 20 locations in northwest Germany. Parasitol Res. 2009;105:351–7.
50. Larska M, Grochowska M, Lechowski L, Żmudziński JF. Abundance and species composition of *Culicoides* spp. biting midges near cattle and horse in south-eastern Poland. Acta Parasitol. 2017;62:739–47.
51. Versteirt V, Balenghien T, Tack W, Wint WA. First estimation of *Culicoides imicola* and *Culicoides obsoletus*/ *Culicoides scoticus* seasonality and abundance in Europe. EFSA Support Publ. 2017;14
52. Meiswinkel R, Elbers ARW. The dying of the light: crepuscular activity in *Culicoides* and impact on light trap efficacy at temperate latitudes. Med Vet Entomol. 2016;30:53–63.
53. Blomberg O, Itämies J, Kuusela K, Itamies J. Insect catches in a blended and a black light-trap in northern Finland. Oikos. 1976;27:57.
54. Carpenter S, Wilson A, Barber J, Veronesi E, Mellor P, Venter G, et al. Temperature dependence of the extrinsic incubation period of orbiviruses in *Culicoides* biting midges. PLoS One. 2011;6:e27987.
55. Harrup LE, Bellis GA, Balenghien T, Garros C. *Culicoides* Latreille (Diptera: Ceratopogonidae) taxonomy: current challenges and future directions. Infect Genet Evol. 2015;30:249–66.
56. Purse B, Nedelchev N, Georgiev G, Veleva E, Boorman J, Denison E, et al. Spatial and temporal distribution of bluetongue and its *Culicoides* vectors in Bulgaria. Med Vet Entomol. 2006;20:335–44.
57. Hartemink NA, Purse BV, Meiswinkel R, Brown HE, de Koeijer A, Elbers ARW, et al. Mapping the basic reproduction number (R0) for vector-borne diseases: a case study on bluetongue virus. Epidemics. 2009;1:153–61.
58. Searle KR, Blackwell A, Falconer D, Sullivan M, Butler A, Purse BV. Identifying environmental drivers of insect phenology across space and time: *Culicoides* in Scotland as a case study. Bull Entomol Res. 2013;103:155–70.
59. Takken W, Verhulst N, Scholte E, Jacobs F, Jongema Y, van Lammeren R. The phenology and population dynamics of *Culicoides* spp. in different ecosystems in the Netherlands. Prev Vet Med. 2008;87:41–54.
60. Carpenter S, Lunt HL, Arav D, Venter GJ, Mellor PS. Oral susceptibility to bluetongue virus of *Culicoides* (Diptera: Ceratopogonidae) from the United Kingdom. J Med Entomol. 2006;43:73–8.
61. Balenghien T, Pagès N, Goffredo M, Carpenter S, Augot D, Jacquier E, et al. The emergence of Schmallenberg virus across *Culicoides* communities and ecosystems in Europe. Prev Vet Med. 2014;116:360–9.
62. Ségard A, Gardès L, Jacquier E, Grillet C, Mathieu B, Rakotoarivony I, et al. Schmallenberg virus in *Culicoides* Latreille (Diptera: Ceratopogonidae) populations in France during 2011–2012 outbreak. Transbound Emerg Dis. 2017;65(1):e94-e103.

3.2 Manuscript II

Monthly variation in the probability of presence of adult *Culicoides* populations in nine European countries and the implications for targeted surveillance

Ana Carolina Cuéllar, Lene Jung Kjær, Andreas Baum, Anders Stockmarr, Henrik Skovgard, Søren Achim Nielsen, Mats Gunnar Andersson, Anders Lindström, Jan Chirico, Renke Lühken, Sonja Steinke, Ellen Kiel, Jörn Gethmann, Franz J. Conraths, Magdalena Larska, Marcin Smreczak, Anna Orłowska, Inger Hamnes, Ståle Sviland, Petter Hopp, Katharina Brugger, Franz Rubel, Thomas Balenghien, Claire Garros, Ignace Rakotoarivony, Xavier Allène, Jonathan Lhoir, David Chavernac, Jean-Claude Delécolle, Bruno Mathieu, Delphine Delécolle, Marie-Laure Setier-Rio, Roger Venail, Bethsabée Scheid, Miguel Ángel Miranda Chueca, Carlos Barceló, Javier Lucientes, Rosa Estrada, Alexander Mathis, Wesley Tack and Rene Bødker.

Manuscript accepted for publication in *Parasites & Vectors*.

Monthly variation in the probability of presence of adult *Culicoides* populations in nine European countries and the implications for targeted surveillance

Ana Carolina Cuéllar[1*], Lene Jung Kjær[1], Andreas Baum[2], Anders Stockmarr[2], Henrik Skovgard[3], Søren Achim Nielsen[4], Mats Gunnar Andersson[5], Anders Lindström[5], Jan Chirico[5], Renke Lühken[6], Sonja Steinke[7], Ellen Kiel[7], Jörn Gethmann[8], Franz J. Conraths[8], Magdalena Larska[9], Marcin Smreczak[9], Anna Orłowska[9], Inger Hamnes[10] Ståle Sviland[10], Petter Hopp[10], Katharina Brugger[11], Franz Rubel[11], Thomas Balenghien[12], Claire Garros[12], Ignace Rakotoarivony[12], Xavier Allène[12], Jonathan Lhoir[12], David Chavernac[12], Jean-Claude Delécolle[13], Bruno Mathieu[13], Delphine Delécolle[13], Marie-Laure Setier-Rio[14], Roger Venail[14,18], Bethsabée Scheid[14], Miguel Ángel Miranda Chueca[15], Carlos Barceló[15], Javier Lucientes[16], Rosa Estrada[16], Alexander Mathis[17], Wesley Tack[18] and René Bødker[1]

[1]Division for Diagnostics and Scientific Advice, National Veterinary Institute, Technical University of Denmark (DTU), Lyngby, Denmark [2]Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Lyngby, Denmark [3]Department of Agroecology - Entomology and Plant Pathology, Aarhus University, Aarhus, Denmark [4]Department of Science and Environment, Roskilde University, Roskilde, Denmark [5]National Veterinary Institute (SVA), Uppsala, Sweden [6]Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Reference and Research National Reference Centre for Tropical Infectious Diseases, Hamburg, Germany [7]Department of Biology and Environmental Sciences. Carl von Ossietzky University, Oldenburg, Germany [8]Institute of Epidemiology, Friedrich Loeffler Institute, Greifswald, Germany [9]Department of Virology, National Veterinary Research Institute, Pulawy, Poland [10]Norwegian Veterinary Institute, Oslo, Norway 11Institute for Veterinary Public Health, Vetmeduni, Vienna, Austria [12]CIRAD, UMR ASTRE, F-34398 Montpellier, France [13]Institute of parasitology and tropical pathology of Strasbourg, EA7292, Université de Strasbourg, Strasbourg, France [14]EID Méditerranée, Montpellier, France [15]Laboratory of Zoology, University of the Balearic Islands, Palma, Spain [16]Department of Animal Pathology, University of Zaragoza, Zaragoza, Spain [17]Institute of Parasitology, University of Zürich, Zürich, Switzerland, [18]Avia-GIS NV, Zoersel, Belgium.

*Correspondence: anacu@vet.dtu.dk

Abstract

Background: Biting midges of the genus *Culicoides* (Diptera: Ceratopogonidae) are small hematophagous insects responsible for the transmission of bluetongue virus, Schmallenberg virus and African horse sickness virus to wild and domestic ruminants and equids. Outbreaks of these viruses have caused economic damage within the European Union. The spatio-temporal distribution of biting midges is a key factor in identifying areas with the potential for disease spread. The aim of this study was to identify and map areas of neglectable adult activity for each month in an average year. Average monthly risk maps can be used as a tool when allocating resources for surveillance and control programs within Europe.

Methods: We modelled the occurrence of *C. imicola* and the Obsoletus and Pulicaris ensembles using existing entomological surveillance data from Spain, France, Germany, Switzerland, Austria, Denmark, Sweden, Norway and Poland. The monthly probability of each vector species and ensembles being present in Europe based on climatic and environmental input variables was estimated with the machine learning technique Random Forest. Subsequently, the monthly probability was classified into three classes: Absence, Presence and Uncertain status. These three classes are useful for mapping areas of no risk, areas of high-risk targeted for animal movement restrictions, and areas with an uncertain status that need active entomological surveillance to determine whether or not vectors are present.

Results: The distribution of *Culicoides* species ensembles were in agreement with their previously reported distribution in Europe. The Random Forest models were very accurate in predicting the probability of presence for *C. imicola* (mean AUC = 0.95), less accurate for the Obsoletus ensemble (mean AUC = 0.84), while the lowest accuracy was found for the Pulicaris ensemble (mean AUC = 0.71). The most important environmental variables in the models were related to temperature and precipitation for all three groups.

Conclusions: The duration periods with low or null adult activity can be derived from the associated monthly distribution maps, and it was also possible to identify and map areas with uncertain predictions. In the absence of ongoing vector surveillance, these

maps can be used by veterinary authorities to classify areas as likely vector-free or as likely risk areas from southern Spain to northern Sweden with acceptable precision. The maps can also focus costly entomological surveillance to seasons and areas where the predictions and vector-free status remain uncertain.

Keywords: *Culicoides,* Random Forest, Machine Learning, Europe, monthly distribution, spatial distribution, Presence-Absence data, targeted surveillance

Background

Culicoides (Diptera: Ceratopogonidae) biting midges are small blood-sucking insects responsible for the transmission of viruses causing the European outbreaks of bluetongue (BT) and Schmallenberg diseases in wild and domestic ruminant livestock [1, 2], and for African horse sickness in equids [1, 3]. BTV historically made sporadic incursions into some countries of the Mediterranean Basin (Portugal, Spain, the Greek islands close to Turkey and Cyprus) but from 1998 onwards the situation worsened when five other serotypes spread within France (Corsica), Italy, Greece and countries in the Balkans region [4]. BT was never reported in northern Europe until August 2006, when an unprecedented bluetongue virus (BTV) serotype 8 outbreak started in the border region of Germany, Belgium and the Netherlands and, over the next two years, it spread further over central and northern Europe [5–8]. This epidemic had a significant economic impact within the European Union, as a consequence of the restriction of animal movements and the large amount of financial resources invested in vaccination campaigns and vector surveillance programs [9–11]. In northern Europe, the Afro-Asian vector *Culicoides imicola* Kieffer is absent and therefore, the vector species incriminated in the transmission of BTV were the Palaearctic species belonging to the Obsoletus ensemble *Culicoides obsoletus* (Meigen)/*Culicoides scoticus* Downes & Kettle [12, 13], *Culicoides chiopterus* (Meigen) [14, 15] and *Culicoides dewulfi* Goetghebuer [16]

Many factors contribute to the transmission of vector-borne diseases, including the presence of infected hosts, competent vectors and suitable environmental

temperatures for the pathogen to replicate inside the vector [17]. In the absence of ongoing entomological surveillance, a temporal map of the potential distribution of the vectors is key for health authorities to quickly delimitate possible areas and time periods of risk for disease transmission in the case of an outbreak of a known or emerging vector-borne disease [18–20]. The spatial distribution and phenology of vectors can be predicted from climate and environmental variables such as temperature, precipitation and land cover [18]. Temporal occurrence data (the presence or absence of a species at a specific time) in non-sampled areas or periods can be modelled using statistical techniques. This methodology is used to generate species distribution maps depicting the probability of the species being present at a given time [21], thus identifying areas with low or null adult activity and therefore, periods during which animal movements are safe.

Since the start of the BT outbreaks, European authorities have established a series of regulations for BT surveillance including vector monitoring to analyse the seasonal fluctuation of the vector populations and determine the seasonal vector-free periods (SVFP) for different regions [22, 23]. The EU defines SVFP by using a threshold on the abundance of female specimens, considering the parity stage of the *Culicoides* caught in the traps. This approach has been used to estimate the SVPF in Scotland for species of the Obsoletus group [24]. The authors estimated phenological events for each species such as the start and end of the SVFP. Brugger et al. [23] estimated vector-free periods in Austria using an approach based on the European Commission definition but without considering parity stage of female specimens. In the present study, we identified months where adult activity is null or very low, based on the monthly mean abundance for each farm, without considering the parity of the specimens collected as previously proposed by the EU legislation. Our definition of adult activity is different but comparable to the vector-free season defined by this legislation and, therefore, we keep the term “vector-free” season or period to refer to a period of the year with neglectable adult activity.

The SVFP during the winter was not ubiquitous across all European countries. Austria [23], Switzerland [25] and Sweden [26] reported the existence of a SVFP, while

other countries such as Germany, France, Belgium and the Netherlands reported that a SVFP might not exist in these countries [16, 27–29]. Imposing restrictions of animal movement in areas where the vector is not present has a negative economic impact as the restriction is unnecessary. On the other hand, allowing animal movement in areas where the vector is present poses a risk of spreading infections to new areas, if environmental conditions are suitable for the virus to develop inside the vector. Being able to define vector-free areas and periods is not only useful for BT management, but also for emerging *Culicoides*-borne diseases in the future. For instance, Schmallenberg virus appeared suddenly in 2011 in Germany, and spread throughout 29 European countries [30], causing economic losses for sheep and cattle farmers [31]. In addition, the spread of African horse sickness has previously been reported in horses in Spain in 1966 and Spain and Portugal from 1987 to 1990 [32]. Knowing the geographical distribution of vectors allows veterinary authorities to focus control measurements in those areas at a specific time of year.

In this study, we used entomological data of *C. imicola*, Obsoletus ensemble and Pulicaris ensemble collected from nine European countries over a seven-year period. This entomological dataset was used previously to analyse the temporal fluctuation at different latitude bands for Europe, to analyse the start of the season at the geographical NUTS level and to interpolate the observed *Culicoides* abundance spatially [32]. In this work, we use the machine learning algorithm “Random Forest” (RF) to model the average monthly presence/absence observed and predict the probability of presence of *C. imicola*, Obsoletus ensemble and Pulicaris ensemble in unsampled areas, using climatic and environmental variables as predictors. The aim of this work was to predict areas and months likely to be free of biting midges or likely to have vectors as well as areas of uncertain status that need to be targeted for entomological surveillance in case of an outbreak. The resulting maps represent the first spatial distribution model for a transect comprising nine European countries from southern Spain to northern Sweden. The maps are useful tools as inputs for decision making by veterinary authorities to detect areas

with adult activity and use this information to focus financial resources for active entomological surveillance programs.

Methods

Culicoides data

We used entomological data collected in farms from Spain, France, Germany, Switzerland, Austria, Denmark, Sweden, Norway and Poland between 2007 and 2013 as part of national surveillance programs or research projects [33]. For each trap site, observations consisted of the number of *C. imicola*, Obsoletus ensemble [*C. obsoletus*, *C. scoticus*, *Culicoides montanus* Shakirzjanova, *Culicoides chiopterus* (Meigen) and *C. dewulfi*] and Pulicaris ensemble [*Culicoides pulicaris* (Linnaeus) and *Culicoides punctatus* (Meigen)]. *Culicoides* biting midges were sampled from a total of 904 livestock farms comprising 31,429 trap collections. Onderstepoort traps were used for sampling biting midges, except for Germany (Biogents Sentinel traps) and in Spain (mini CDC traps). For these two countries, we multiplied the number of *Culicoides* for each observation by a conversion factor to make the number of specimens comparable between the different trapping methods. Details of both the sample protocols and the conversion factors used have been published previously [33].

For *C. imicola* and each of the *Culicoides* ensembles, we split the observation data set into 12 subsets according to month of the year. For each 12 monthly dataset, we calculated the average abundance on each farm for each year sampled. This resulted in 12 datasets with farms containing one monthly average abundance per year sampled. Then, we classified each monthly average each year into Presence or Absence according to the average abundance of the vector. Based on the European Union regulation [22] for the definition of the SVFP, in which an abundance threshold of biting midges is proposed to define Presence or Absence, we considered each monthly average for each year as Presence when it was above or equal to an abundance threshold of five midges for the Obsoletus and Pulicaris ensembles, and one specimen for *C. imicola*. Even

though the European Union definition of Presence is based on the catch of five parous specimens per observation, we here considered the number of midges without differentiating females into their gonotrophic stage because this information was missing for some of the countries. This will result in a more conservative definition of SVFP. Our approach also differed from the approach used by the EU commission as for each farm we only classified the monthly average each year into Presence or Absence, and not each of the individual observations (when there were several observations per month).

We constructed preliminary Random Forest (RF) models using occurrence data from January and February. The data collected in this period did not include any farms from northern Scandinavia. The resulting models predicted the occurrence of biting midges in January and February in this region (data not shown). However, earlier studies have reported an absence of biting midges in the Scandinavian peninsula during winter [26, 34]. Therefore, it was useful to provide pseudo-absence points to the models in order to increase their accuracy for predicting absences in the area. For January and February, we created 11 random pseudo-absence points above 60 degrees latitude in the highlands in Norway, central and northern Sweden and Finland and were added by hand using ArcMap 10.1 (ESRI, Redlands, CA, USA) (Fig. 1).

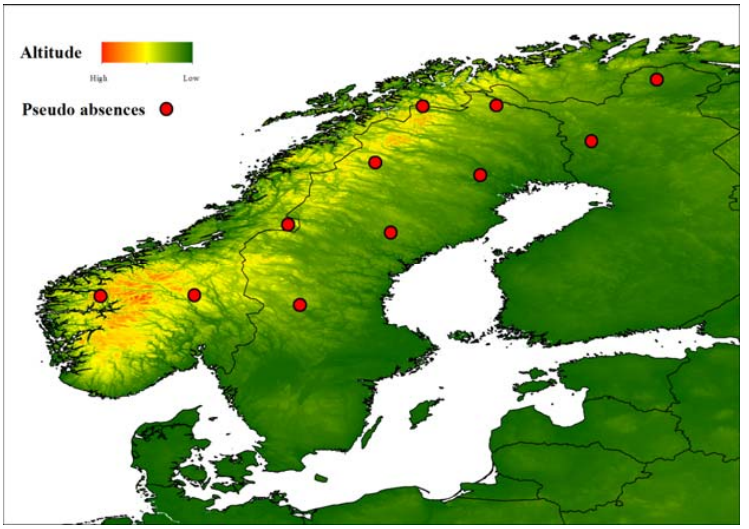


Fig. 1 Eleven pseudo-absence points added to Norway, Sweden and Finland for January and February

Predictor variables

We used raster files (images) of 112 environmental and climatic variables, land cover and livestock density, each with a 1 km^2 spatial resolution.

The environmental predictors included Mid-infrared (MIR), daytime Land Surface Temperature (dLST), nighttime Land Surface Temperature (nLST), Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) as predictor variables. Each variable was derived from a MODIS temporal series from 2001 to 2012, and subjected to Temporal Fourier Analysis (FTA) [35]. For each environmental variable, the resulting products of FTA were the 14 images described in Table 1. This dataset was originally created by the TALA research group at the Department of Zoology at Oxford University, and was provided through the EDENext project [36].

Fourier component	Description
A0	Fourier mean for the entire time series
A1	Amplitude of annual cycle
A2	Amplitude of bi-annual cycle
A3	Amplitude of tri-annual cycle
P1	Phase of annual cycle
P2	Phase of bi-annual cycle
P3	Phase of tri-annual cycle
DA	Proportion of total variance due to all three cycles
D1	Proportion of total variance due to annual cycle
D2	Proportion of total variance due to bi-annual cycle
D3	Proportion of total variance due to tri-annual cycle
MN	Minimum value
MX	Maximum value
VR	Total variance

Table 1 Products of Temporal Fourier Analysis obtained from a single variable. Each product corresponds to a raster image (1 km^2 resolution) derived from a single environmental variable (for instance, NDVI)

We also included WORLDCLIM altitude data (digital elevation model) and bioclimatic variables as climatic predictors for *Culicoides* distribution. BIOCLIM images were obtained from the WORLDCLIM database [37, 38] (Table 2).

Source	**Code**	**Description**
MODIS (Fourier transformed) 2001–2012	MIR	Mid-infrared
	dLST	Daytime land surface temperature
	nLST	Nighttime land surface temperature
	NDVI	Normalized difference vegetation index
	EVI	Enhanced vegetation index
BIOCLIM 1960–1990	BIO 1	Annual mean temperature
	BIO 2	Mean diurnal range: mean of monthly (max. temp - min. temp)
	BIO 3	Isothermality (BIO2/BIO7) (×100)
	BIO 4	Temperature seasonality (standard deviation × 100)
	BIO 5	Max. temperature of warmest month
	BIO 6	Min. temperature of coldest month
	BIO 7	Temperature annual range (BIO5-BIO6)
	BIO 8	Mean temperature of wettest quarter
	BIO 9	Mean temperature of driest quarter
	BIO 10	Mean temperature of warmest quarter
	BIO 11	Mean temperature of coldest quarter
	BIO 12	Annual precipitation
	BIO 13	Precipitation of wettest month
	BIO 14	Precipitation of driest month
	BIO 15	Precipitation seasonality (coefficient of variation)
	BIO 16	Precipitation of wettest quarter
	BIO 17	Precipitation of driest quarter
	BIO 18	Precipitation of warmest quarter
	BIO 19	Precipitation of coldest quarter
	Altitude	Digital elevation model (DEM)
Corine Land Cover[a]	CLC 12	Non-irrigated arable land
	CLC 13	Permanently irrigated land
	CLC 15–17	Vineyards, fruit trees and berry plantations, olive groves
	CLC 18	Pastures

CLC 19	Annual crops associated with permanent crops
CLC 20	Complex cultivation patterns
CLC 21	Land principally occupied by agriculture with significant areas of natural vegetation
CLC 22	Agro-forestry areas
CLC 23	Broad-leaved forest
CLC 24	Coniferous forest
CLC 25	Mixed forest
CLC 26	Natural grasslands
CLC 29	Transitional woodland-shrub
CLC 35	Inland marshes
CLC 40	Water courses
CLC 41	Water bodies

[a]CLC plus the number refers to the CORINE land cover class used for modelling

Table 2 MODIS Fourier-transformed, BIOCLIM and Corine Land Cover predictors used to model the probability of *Culicoides* presence

We used a Corine Land Cover (CLC) map with 250 m pixel resolution to extract information on 16 relevant land cover classes (Table 2). For each class, we created a binary image with pixel values of 1 and 0 according the presence or absence of the class. Due to the higher spatial resolution of the CLC map compared to the other predictors, we resampled each of the binary class images to a resolution of 1 km^2. This was done by overlaying a grid with cells of 1 km^2 resolution. To each of these cells, we assigned the sum of all pixels with a value of 1 within them. Each 1 km^2 cell of the grid was made up of 16 (4 × 4) pixels of the original CLC map. This resulted in new images for each land cover class with a pixel resolution of 1 km^2, representing the frequency of each of the 16 different classes found in every 1 km^2 area (pixel) on a scale of 0–16. CLC map was obtained from the European Environment Agency website [39].

We obtained livestock density data for cattle, goats, sheep, small ruminants and chickens from the Food and Agriculture Organization repository “GeoNetwork”. This dataset consisted of a series of raster files with information regarding livestock density at a global scale (“The gridded livestock of the world”) [40].

Modelling the probability of presence

Combining our *Culicoides* data with the predictors, we explored modelling approaches using VECMAP© software, v.2.0.16350.2473. For the final modelling of each month and each species, we used the Random Forest (RF) machine learning technique [41, 42] in R v.3.4.2 [43] (packages *caret* [44] and *randomForest* [45]) to model the probability of presence (PP) in the nine European countries using the Presence/Absence observations calculated at each farm. For each month we obtained a map showing the PP at the same resolution as the predictors (1 km^2). The RF algorithm consists of an ensemble of decision trees used to predict the probability of class membership where the response variable is categorical (e.g. classification into presence and absence). An advantage of RF is the model's capability of detecting nonlinear relationships between the response and the predictor variables [46] and that RF can handle a large number of predictor variables [46]. In addition, RF can produce a list of the most important predictors and scale them from 0–100 according to their importance as calculated by permuting each predictor and measuring the prediction error after the permutation [44].

The number of farms sampled varied from month to month. As expected, during summer more farms were sampled compared to winter, as in many countries of northern Europe entomological surveillance is not carried out during the cold winter months. For each monthly dataset, we used a stratified random split to divide the data into two subsets: one included 70% of the farms containing at least one year classified as presence together with the farms with only absence observations (training set). The second subset contained the remaining 30% of the farms as a test set to evaluate model performance [42, 47, 48]. We conducted a stratified random split based on farm ID in order to avoid having observations belonging to the same farm in both the training and the evaluation datasets (Table 3).

Month	Total no. of sampled farms	Training set (70%)	Test set (30%)
January	444	310	134
February	457	319	138
March	473	331	142
April	522	364	158
May	527	368	159
June	518	362	156
July	581	406	175
August	636	445	191
September	620	433	187
October	522	365	157
November	500	349	151
December	448	313	135

Table 3 Total number of farms sampled each month and number of farms in the training and test sets. All observations belonging to a single farm were included in either the training or test set, but never in both

The number of *Culicoides* caught per farm highly varied between the different years. In this work, we considered each farm's monthly classification into Presence or Absence for each year and included them in the training set as independent observations. Therefore, a farm might contain Presence and Absence observations from different years depending on the variation in mean monthly abundance between the different years.

The monthly Presence/Absence data were highly imbalanced, meaning that it contained a high proportion of one of the classes (Presence or Absence), i.e. the majority class. We investigated and compared five different balancing methods (no balancing, down-sampling, oversampling, ROSE[49], SMOTE [50], Tomek [50]) to cope with the imbalance and to improve model performance. We ran cross-validation (CV) for each balancing method 10 times with different random seeds and the best method was chosen according to highest AUC (data not shown). The balancing method chosen to balance the training set was oversampling, which entails duplicating the observations for the

minority class in order to reach the same number of observations as the majority class [42]. We used the balanced training set of each month to train the RF model, and used the test sets to calculate the receiver operating characteristics (ROC) curve [42, 51, 52] and the area under this curve (AUC). We used the AUC as a measurement of model performance. AUC values close to 0.5 indicate that the model is not able to classify new samples better than random, values between 0.7 and 0.8 indicate acceptable model performance, values from 0.8 to 0.9 indicate excellent performance and values above 0.9 are considered outstanding [53]. For each month, we performed 5-fold CV to optimize the model parameter "mtry" (i.e. number of predictors used at each split). The "ntrees" parameter (number of trees of the forest) was set to 1000 trees in all cases.

For *C. imicola*, after the test set was created, we removed all the observations from farms not belonging to Spain or France, as the vector was not found in the seven remaining countries [33]. This reduced the large amount of Absence observations in the test set, which have an influence in the distribution of the classes.

Classification

Classification of predicted probabilities into Presence/Absence classes can be determined using a predetermined threshold (in ecology studies, normally the default is a PP of 0.5 [54]). Here, we were interested in defining a data-dependent threshold, as a predefined threshold of 0.5 might not be optimal [54]. The monthly PP maps obtained from our RF models were classified into three categories. We calculated a lower and upper threshold and all areas with a PP below the lower threshold were considered to be in the Absence class, while the areas with a PP above the upper threshold were classified as Presence areas. Regions with a PP between the two thresholds could not be classified as either Absence or Presence class, and were therefore classified as an Uncertain status category that may be targeted for active vector surveillance. The Absence and Presence classes refer here to the occurrence of adult activity and not to the ecological establishment of the vector, as in the classical species distribution modelling.

Lower and upper thresholds were calculated using the density function for the PP predicted by the model for each test set class (true presence/absence). To define the two thresholds for each month, we derived two gain functions $G_{presence}$, $G_{absence}$ for 100 possible thresholds from 0 to 1, based on the area under the density function for Presence and Absence, respectively. We calculated $G_{presence}$ as the probability of a true presence and subtracted the probability of a misclassified presence multiplied by a parameter δ, which indicates the cost of a misclassified presence relative to a true presence. Similarly, we calculated $G_{absence}$ as the probability of a corrected classified absence (true absence) and subtracted the probability of misclassified absence multiplied by parameter γ, which indicates the cost of a misclassified absence relative to a true absence. Setting $\delta = 2$, for example, means that the cost of a false positive classification is twice the gain of a true positive classification. The gain value can be considered in terms of timely initiation of countermeasures and a lower probability of an epidemic and trade restrictions, while the loss value would be the cost to the farm and society of incorrectly applied countermeasures. Similarly, for the interpretation of γ, the gain of a true negative classification and the loss from a false negative classification can be likened to being declared free from disease, with the cost to both farmer and society of a subsequent discovery of the disease. Similar considerations can be used to relate δ and γ to each other. If, for example, we set $\delta = \rho * \gamma$ in Equation 1, the cost of misclassifying a presence is ρ times the cost of misclassifying an absence. We assign $\delta = 2 * \gamma$ in order to assign twice the importance to the Presence misclassifications compared to Absence misclassifications and we set $\gamma = 2$ to still give some importance to the Absences misclassifications.

The equations for $G_{presence}$, $G_{absence}$ were:

$$G_{presence}(q) = \int_q^1 Presence\ (x)\ dx - \ \delta * \ \int_q^1 Absence(x)\ dx$$

(Eqn. 1)

$$G_{absence}(q) = \int_0^q Absence\ (x)\ dx - \ \gamma * \ \int_0^q Presence(x)\ dx$$

(Eqn. 2)

where q represents the possible threshold value between 0 to 1, and where δ and γ are loss parameters.

To calculate the lower threshold, we used Equation 1 to find the optimal upper threshold when assuming a loss parameter of δ = 4 by optimizing the gain $G_{presence}$. Similarly, Equation 2 was used to find the optimal lower threshold, assuming a loss parameter γ = 2. The upper and lower thresholds depend on the predictive power of the model, being more separated when the overlapping between classes is large. If the model performance is good, the overlapping between classes will be less and the two thresholds will be closer together.

In order to evaluate the sensitivity of the thresholds to the distribution of different test sets, we divided each monthly test set into ten equally sized folds (10 subsets) and calculated the density functions using nine out of the ten folds. This procedure was repeated for all the different folds (10 times), excluding a different fold each time, and plotted the new lower and upper thresholds together in the same graph. We applied this 10-fold cross-validation scheme to compare the threshold calculated with different subsets of the test set versus the thresholds calculated using all the observations of the test set.

We classified the monthly probability maps into the three classes: “Absence”, “Uncertain” and “Presence” using the thresholds calculated from all the observations of the test set.

Results

Obsoletus ensemble

The 12 models were shown to perform well for the Obsoletus ensemble, with an AUC ranging from 0.76 in June and December to 0.91 in November (mean AUC = 0.84) (Fig. 2).

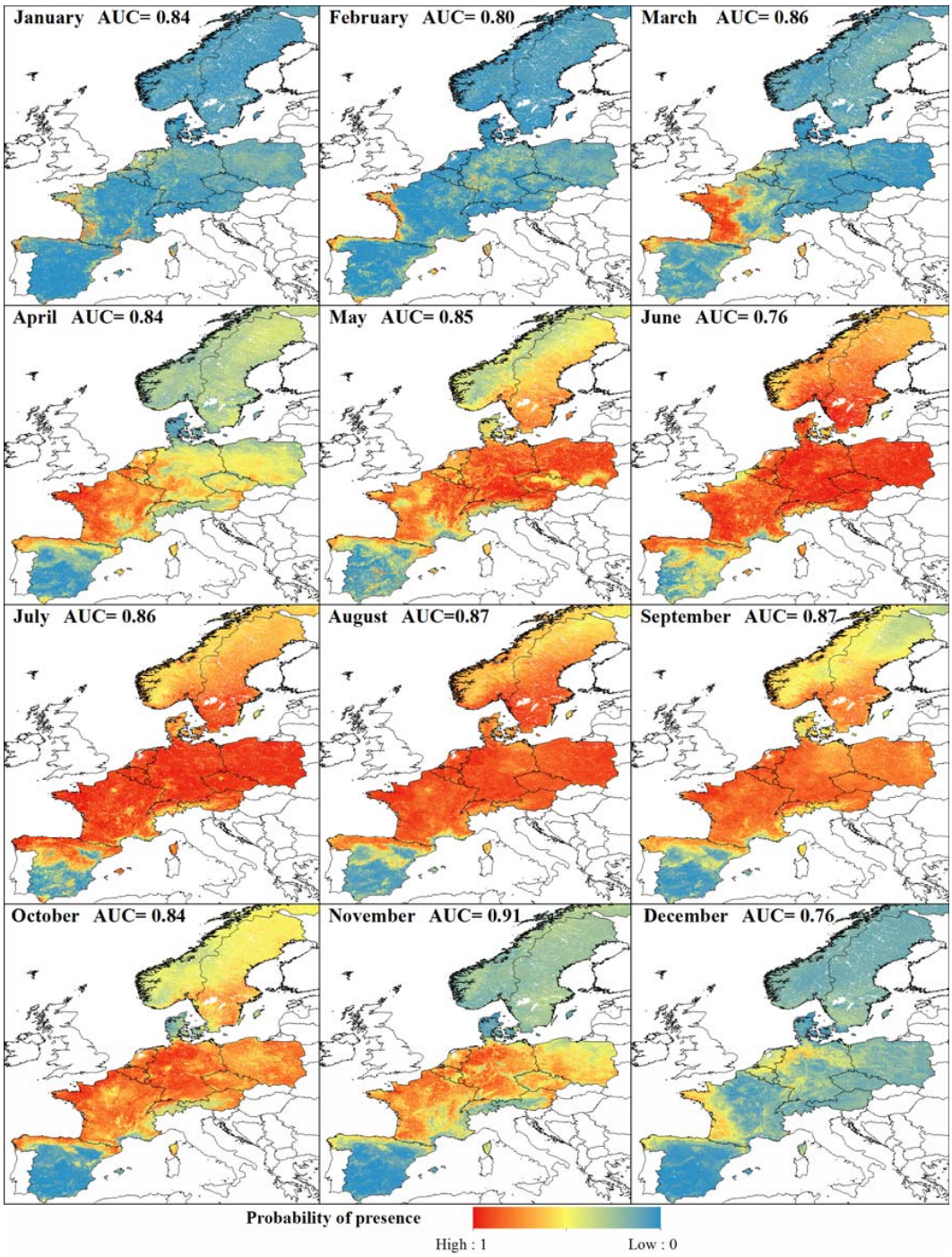


Fig. 2 Predicted monthly probability of presence of Obsoletus ensemble. Monthly model performance is shown as the AUC value

The majority class shifted from Absence in December-March, to Presence in April- November, and the models generally had good predictive power when predicting the majority class. However, the models performed less well when predicting the minority class. For January and February, the model predicted the Presence class relatively poorly, with a relatively flat density function (Fig. 3). The additional thresholds calculated using 10-fold CV were similar to the main threshold, indicating that the distribution of classes in the test set were robust when subtracting 10% of the data. The lower thresholds showed more variation compared to the variation of the upper thresholds (Fig. 3).

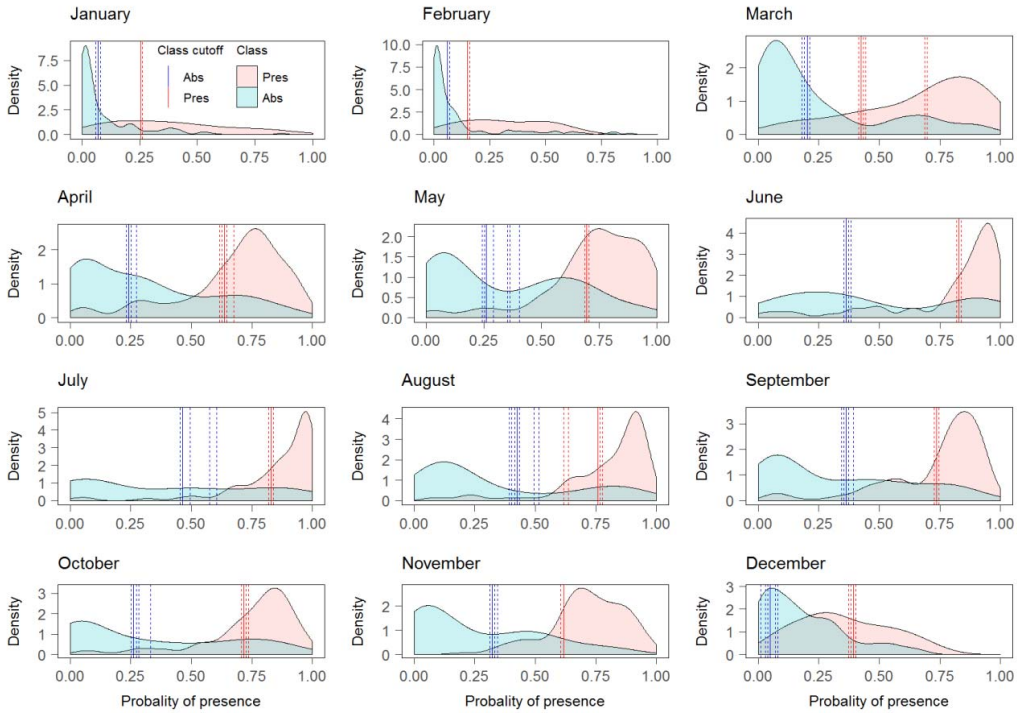


Fig. 3 Obsoletus ensemble: monthly distribution of Presence and Absence classes of the test set samples as a function of their predicted probability of presence. Dashed lines show the additional thresholds calculated from 10-fold CV

Classifications did not result in clearly delineated geographical zones for the three classes (Presence, Absence and Uncertain), although spatial patterns were observed (Fig. 4). In January, the Obsoletus ensemble was predicted present in areas within the western part of France, northern coast of Spain and in scattered areas of Germany, and it was predicted absent from northern and central Scandinavia, eastern France and parts of Germany. The Uncertain class area was present in southern Scandinavia, eastern Germany and Poland. In February the Presence area in western France and the northern coast of Spain appeared clearly segregated while more dispersed patches appeared in Germany and Poland. The Uncertain class area was reduced to patches in Germany, Poland and a small portion of southern Sweden. During March, the Presence area extended further west into France, while the Absence area was clearly concentrated in the eastern part of Europe and Scandinavia. The Uncertain area was a more coherent intermediate region between these two areas, found in eastern France, Belgium and the Netherlands. In April, the Presence class expanded from western France occupying most of France while the eastern part of the study area and Scandinavia remained in the Uncertain area. From May onwards, the general pattern showed the Obsoletus ensemble to be widely distributed in France, Germany, Austria, Switzerland, Poland and southern Scandinavia. The Absence class areas were located in southern Spain during this period. In November, Scandinavia was classified as an Absence class area together with Spain (except the northern coast of Spain, that was included in Presence area). Finally, in December the Presence class was clustered in western France and some patches in northern Germany while the remaining areas, with exception of southern Spain, appeared classified as Uncertain areas, including the Scandinavian peninsula (Fig. 4).

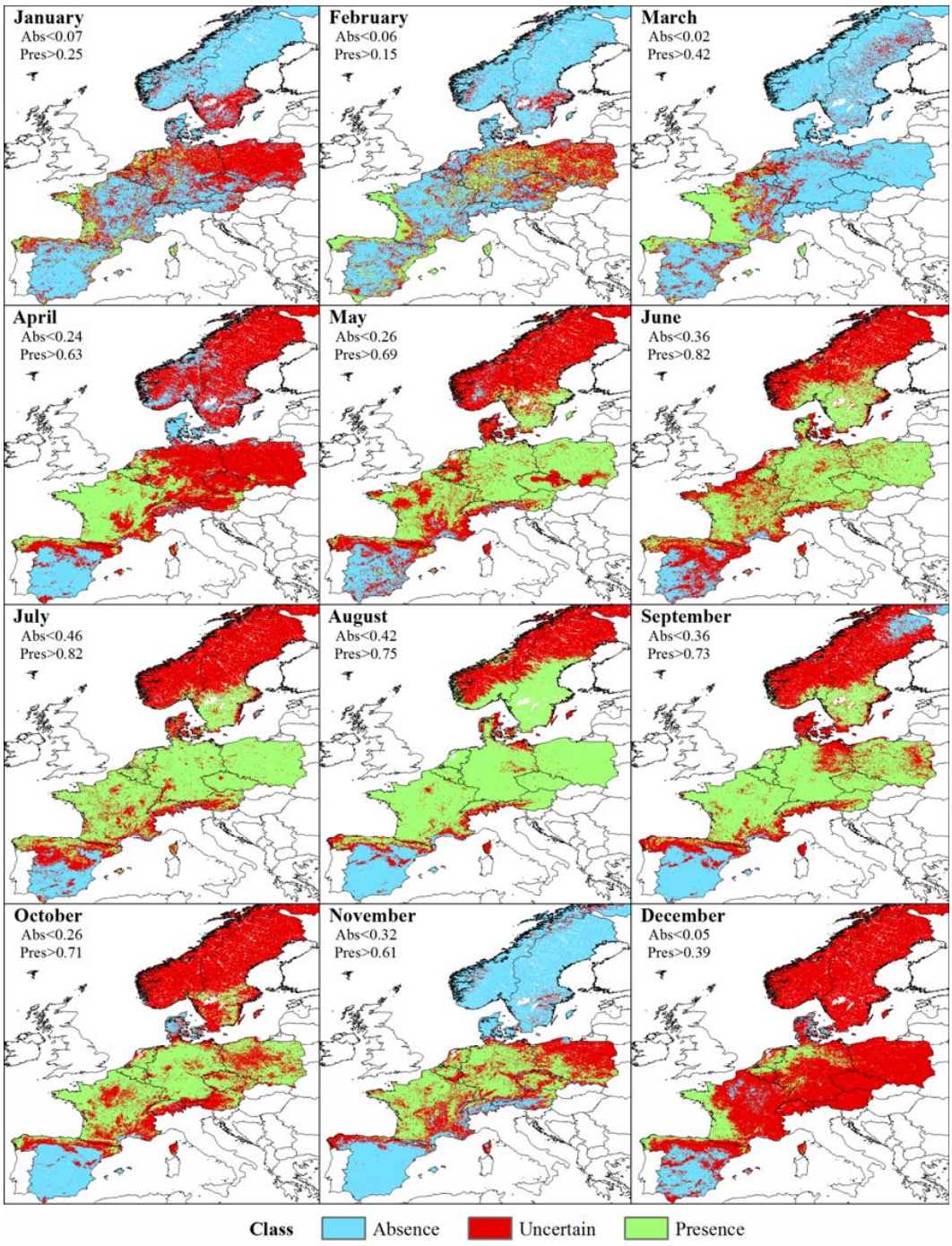


Fig. 4 Classification of the predicted probability of presence of Obsoletus ensemble into Absence, Presence and Uncertain areas at a 1 km^2 resolution

Pulicaris ensemble

The RF models performed less well in predicting the PP for the Pulicaris ensemble. The mean AUC was 0.81, ranging from 0.69 in April to 0.92 in December (Fig. 5).

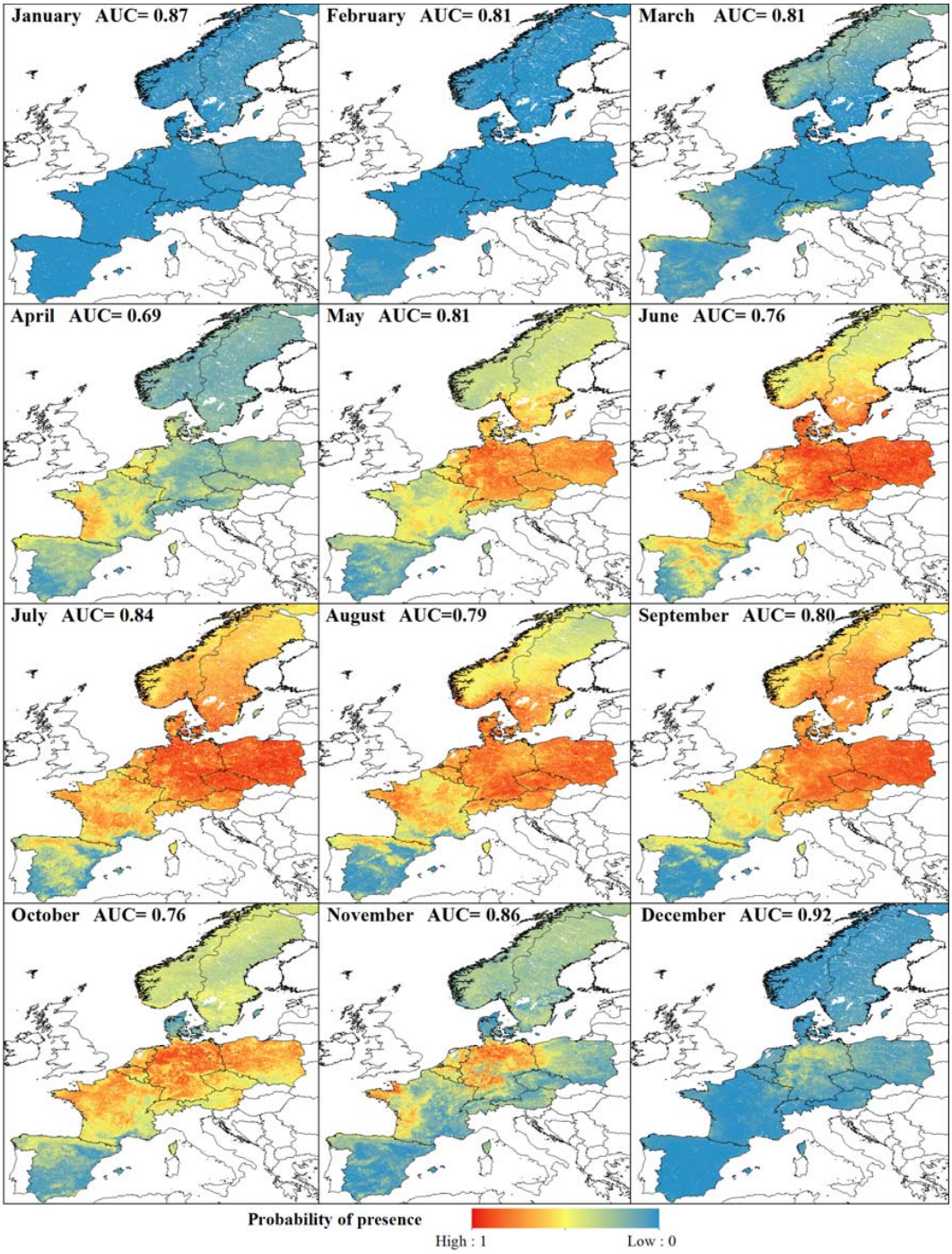


Fig. 5 Predicted monthly probability of presence of Pulicaris ensemble. Monthly model performance is shown as the AUC value

For January, the test set contained only three Presence observations from a single farm and the density function and thresholds could not be calculated. Therefore, the PP map could not be classified into the three classes. For February, the PP predicted for the observed Presences were completely included within the range of the PP predicted for the Absence class, meaning that the model was incapable of distinguishing the Presence class. Nevertheless, because both density functions were computed, the lower and upper thresholds were still calculated. The distribution of predicted Presence and Absence areas for the Pulicaris ensemble test set contained larger overlapping areas between both distributions than for the Obsoletus ensemble, resulting in poorer predictive power for distinguishing between the classes. For the months of April, May and June, the distribution of both classes overlapped so much that the lower threshold was calculated as close to 0 to avoid false negative classifications (Fig. 6). For the Pulicaris ensemble, the additional thresholds calculated using 10-fold CV, were similar to the main threshold for all the months, meaning that the distribution of classes in the test set were robust when subtracting 10% of the data. Both lower and upper thresholds seemed to be robust for the different test sets (Fig. 6).

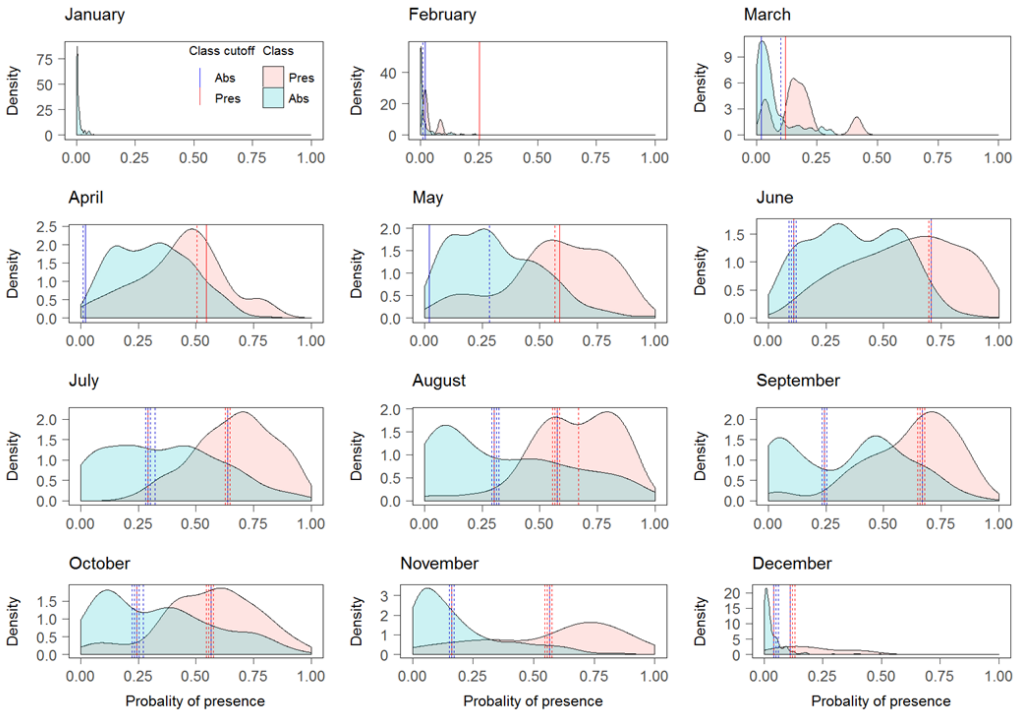


Fig. 6 Pulicaris ensemble: monthly distribution of Presence and Absence classes of the test set samples as a function of their predicted probability of presence. Dashed lines show the additional thresholds calculated from 10-fold CV

Due to the lack of Presence observations in January, we could not define thresholds for classifying the PP map. In February, because PP of the observed Presence observations were completely included in the range of the PP of the Absence class, we decided not to classify the map as the model was incapable of distinguishing the Presence class and would lead to an incorrect interpretation of the classification. In March, the Pulicaris ensemble was predicted to be present on the west coast of France, northern coast of Spain and in central and northern Scandinavia, while the Absence class was distributed in eastern France, Germany and Poland. The Uncertain area was located between the Presence and Absence class. During April, May and June, the model was able to predict the Presence class but it was incapable of distinguishing the Absence class, resulting in classification only for the Presence and Uncertain class. From July to

October, the Presence class extended towards the eastern part of the study area while the Uncertain class occupied northern Scandinavia. During September, the Uncertain class was additionally found in France. In November, the Presence areas were located mostly in Germany and some patches in France while Scandinavia was classified into the Uncertain class. The Absence class was predicted in Denmark and southern Spain. During December, the Absence class was localized in Spain, France and northern Scandinavia while the Presence class remained in some patches in Germany (Fig. 7).

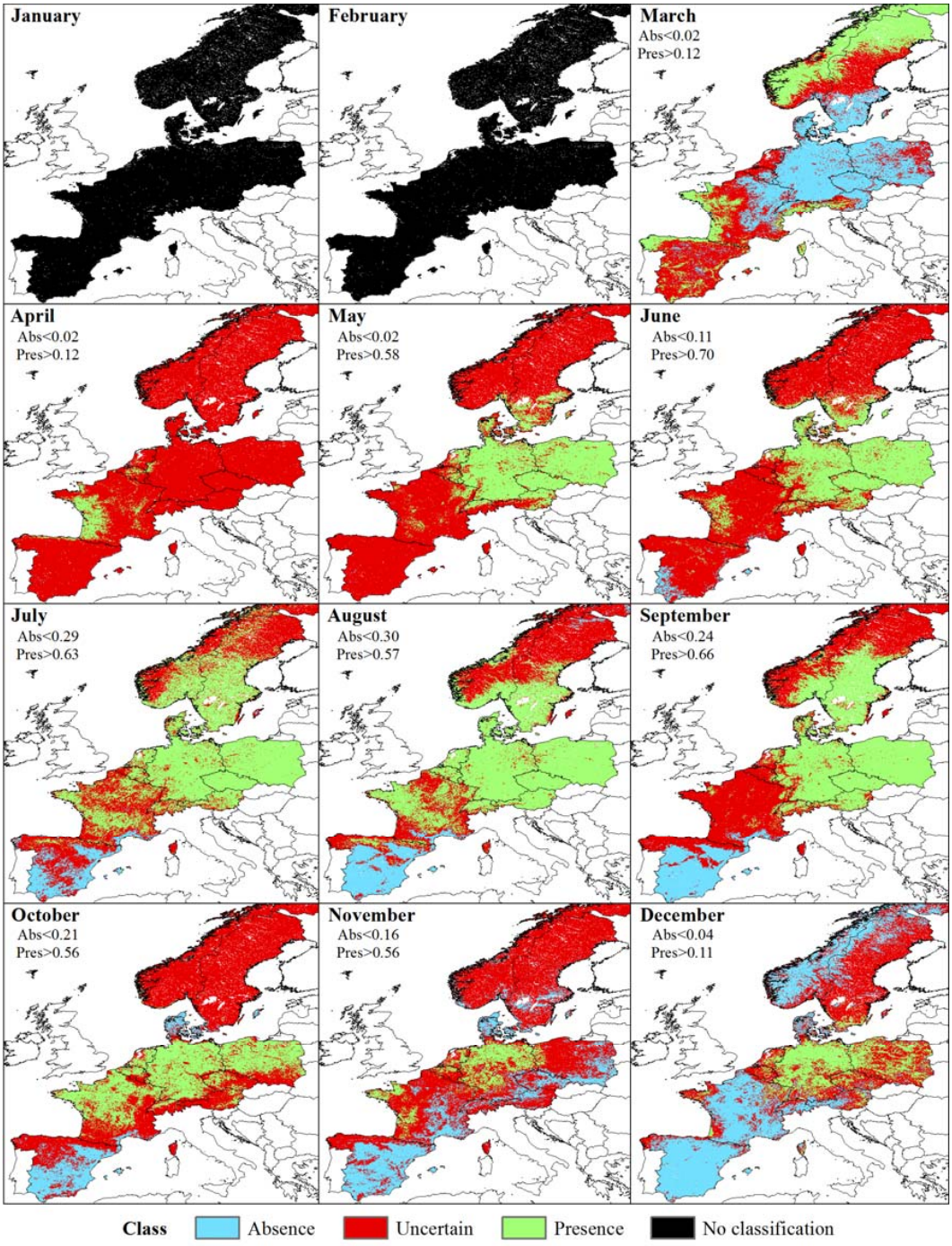


Fig. 7 Classification of the predicted probability of presence of Pulicaris ensemble into Absence, Presence and Uncertain areas at a 1 km^2 resolution

Culicoides imicola

The RF models for *C. imicola* had a very high accuracy for distinguishing the Presence and Absence classes. The models had a mean AUC of 0.95, ranging from 0.92 in January to 0.97 in August (Fig. 8).

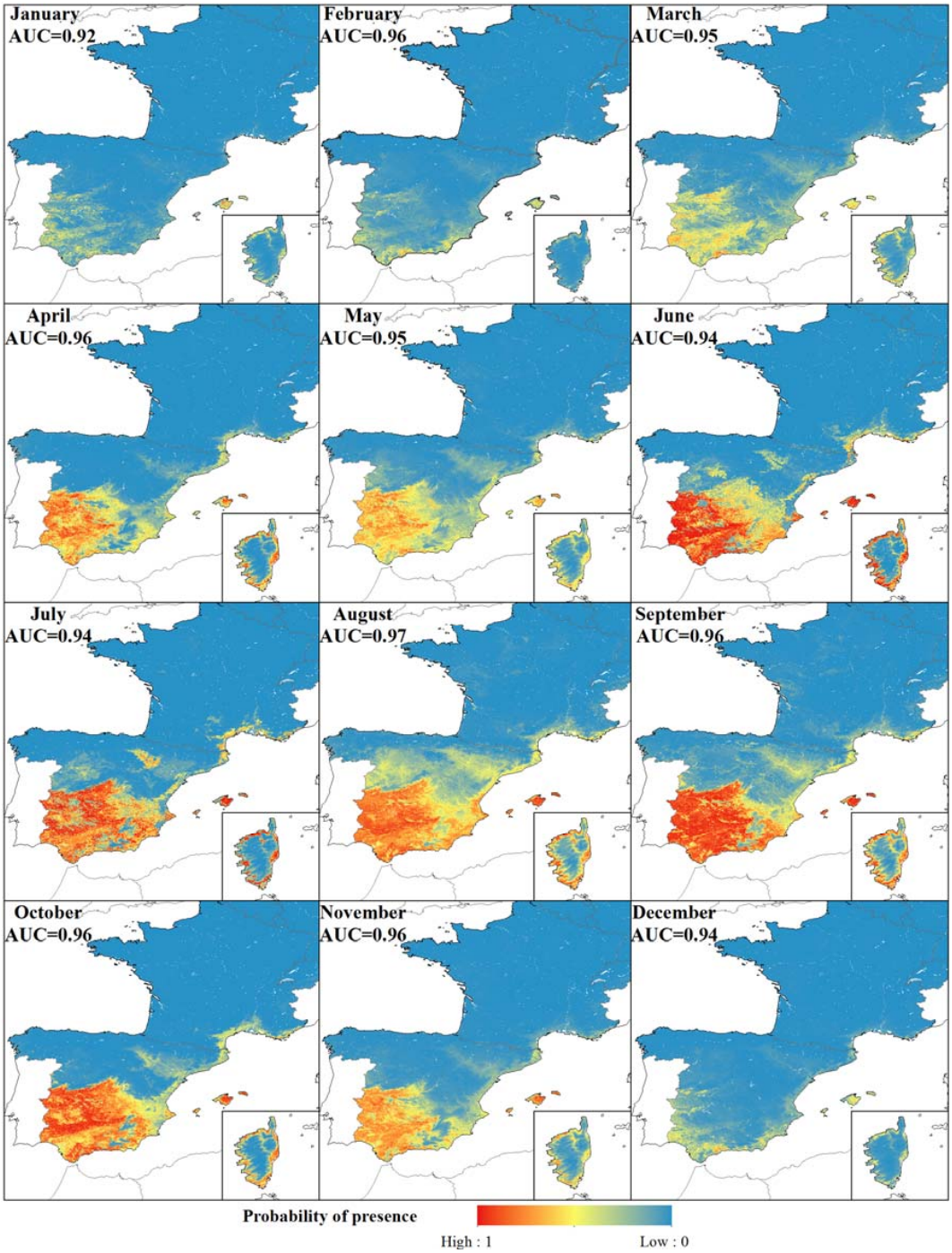


Fig. 8 Predicted monthly probability of presence of *C. imicola*. Monthly model performance is shown as the AUC value

The RF models predicted the *C. imicola* Absence class very well. Absence constituted the majority class for all months as the species was only found in Spain and southern France. The Presence class was less well predicted, as reflected in a flatter distribution. Nevertheless, the model was able to distinguish both classes, resulting in a narrow area of uncertainty between the lower and upper thresholds (Fig. 9). The additional thresholds calculated using 10-fold CV, were similar to the main threshold, indicating that the distribution of classes in the test set were robust when subtracting 10% of the data. The upper thresholds showed more variation compared to the variation in the lower thresholds. Particularly April, July and November seemed to have upper thresholds sensitive to the class distribution of the test set (Fig. 9).

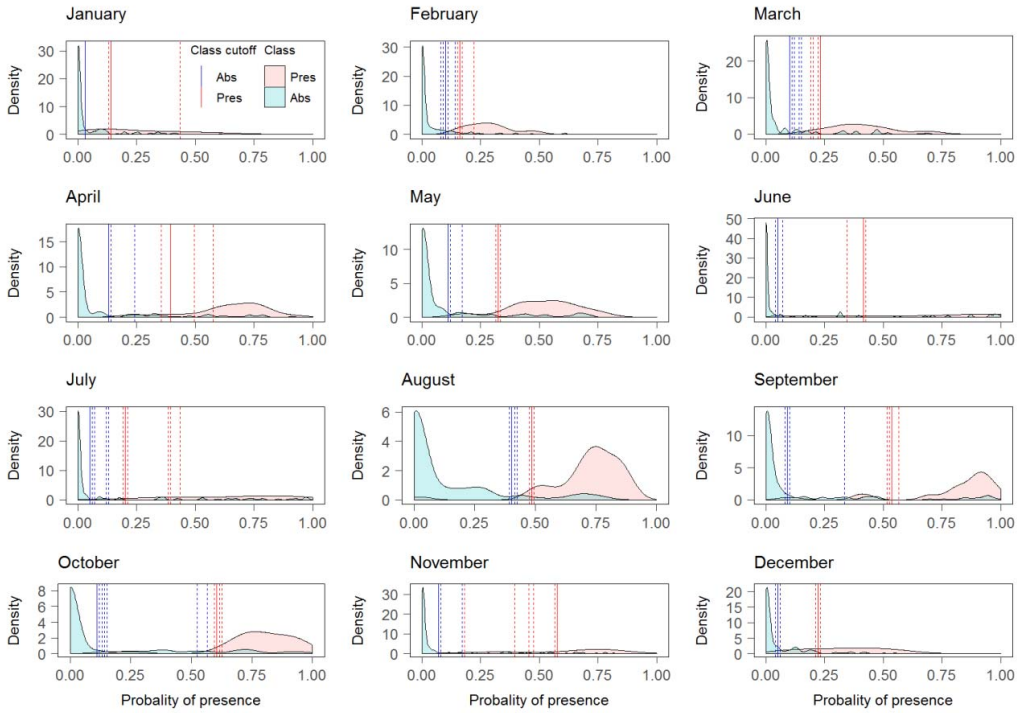


Fig. 9 *Culicoides imicola*: monthly distribution of Presence and Absence classes of the test set samples as a function of their predicted probability of presence. Dashed lines show the additional thresholds calculated from 10-fold CV

Compared to the models for the Obsoletus ensemble, the models for *C. imicola* resulted in a clearer geographical division into three separate coherent zones. *Culicoides imicola* was found to be present in January and February in some areas in southern Spain, the Balearic Islands and Corsica. Uncertain areas were identified in central Spain, while the Absence regions were located in northern Spain and most of France with the exception of the southern coast. From March onwards, the Presence region extended northwards, occupying the southern and central regions of Spain until October, when it retracted back to the southern coast of Spain during late autumn. On Corsica, the Presence areas were located around the coast, with the vector being absent inland. The Uncertain area was always clearly located between the Presence and Absence areas and was generally small due to the high accuracy of the model in distinguishing between Presence and Absence classes (Fig. 10).

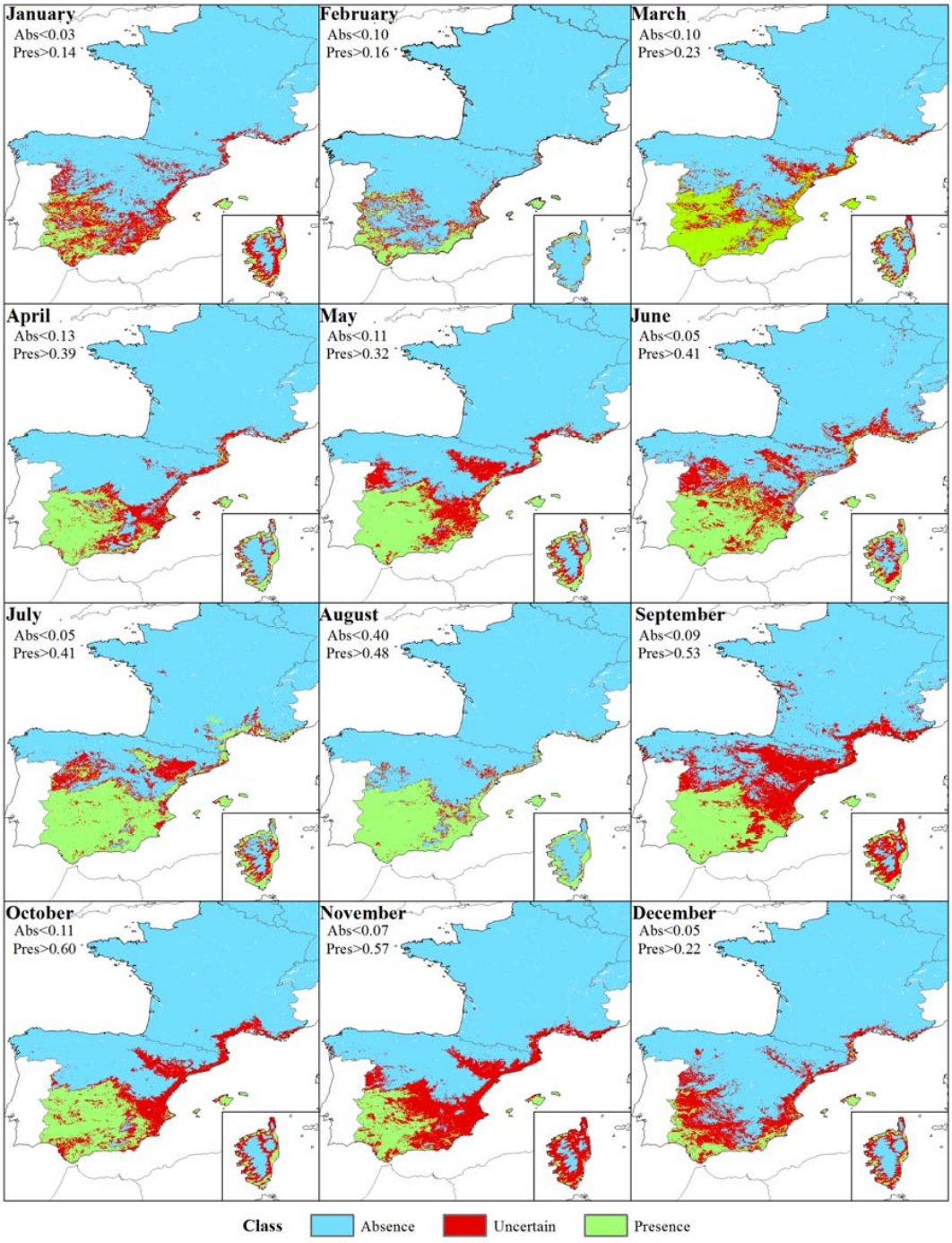


Fig. 10 Classification of the predicted probability of presence of *C. imicola* into Absence, Presence and Uncertain areas at a 1 km^2 resolution

Important predictors

The most important predictors driving the distribution of the Obsoletus ensemble, Pulicaris ensemble and *C. imicola* were related to temperature and precipitation for most months (dLST_MN, nLST_A0, nLST_MX, BIO 10, BIO 18, BIO 5). EVI- and NDVI-derived variables were the most important for some months and for some of the taxa, but with lesser importance compared to temperature and precipitation. Corine land cover classes were not selected as important variables and only one class (CLC 12: non irrigated arable land) was selected for Pulicaris during August. A similar situation occurred for the animal density variables, in which the only variable appearing in the top 5 most important variables was sheep density for the Pulicaris ensemble. Altitude was selected as an important variable only for the Obsoletus and Pulicaris ensembles, for the month of December (Additional file 1).

Discussion

This study was based on the most extensive *Culicoides* dataset created to date. For these prediction maps, we used 31,429 *Culicoides* trap catches from nine European countries from 2007 to 2013 [33]. The objectives of this work were to predict the monthly probability of *Culicoides* presence and to demarcate regions of Europe into three presence classes, each for *C. imicola* and the Obsoletus and Pulicaris ensembles. We also identified areas and periods when the model was not able to predict with reasonable certainty. In these areas, targeted entomological surveillance programs implemented by the CVO's of European Union member states are needed to clarify the present entomological status in case of an outbreak. The maps presented here can be used to determine vector-free areas (Absence areas) and areas where the vector can be found. The Absence and Presence areas were delimitated to minimize misclassification errors, making these classes more accurate in terms of the occurrence of *Culicoides*.

The models generated for the Obsoletus ensemble performed well for all months, and we were able to detect a spatial pattern in the three classes. However, the Absence

and Presence classes were not completely separated by the model, and some geographical areas with Uncertain status were found among the Presence or Absence areas. For some of the months, our RF models were not able to clearly distinguish the minority class from the majority class, resulting in the threshold from the gain function being moved to the extremes to avoid misclassifications. This, in turn, resulted in a large Uncertain area that should potentially be targeted for costly entomological surveillance. This was the case for the Obsoletus ensemble during August, when the vector was indeed present in most of Europe but where our models classified the status as Uncertain in many smaller areas. For instance, in December the model predicted a large Uncertain status area that occupied most of the Scandinavia peninsula while the cold winter conditions make it unlikely that specimens will be found in northern Scandinavia. The Uncertain status areas should be interpreted with care and expert knowledge must be considered when making decisions regarding implementation of surveillance programs. The maps presented here are merely intended as tools and inputs to decision makers for long-term planning and in case of outbreaks in areas without ongoing entomological surveillance. The presented maps are based on a given gain function, but the gain function should reflect the severity of the vector borne diseases with an increasing emphasis on sensitivity as the severity of a disease increases.

In our models, the most important variables for the Obsoletus ensemble were the minimum daytime land surface temperature in January and February, and temperature- and precipitation-related variables (BIO 5 and BIO 14) throughout the rest of the year. Our results are in agreement with the findings of Calvete et al. [55] and Ducheyne et al. [56] who stated that temperature-related variables were the most important for the Obsoletus group distribution in Spain. Additionally, Purse et al [57] found that temperature had an effect in the occurrence of *C. obsoletus* in Italy. The Obsoletus ensemble are Palaearctic species requiring relatively low temperatures and humid climates for optimal development and survival [58, 59]. Temperature plays an important role in *Culicoides* ecology as it determines the seasonal fluctuation of the vector

populations [60, 61], while humidity has been reported to create the optimal conditions for *C. obsoletus* breeding sites (e.g. dung heaps) [62].

To date, maps showing the PP and distribution of the Obsoletus ensemble for the entire Europe are scarce and incomplete. EFSA developed a website displaying distribution maps of *Culicoides* spp. On this site, a map of *C. obsoletus*/*C. scoticus* shows the distribution of this species [63] but the map is lacking information from some countries in Europe. At country level, some studies predicted the probability of Obsoletus group presence based on entomological data collected [56, 64–66]. Therefore, there is a need for predictions on a continental scale summarizing historical surveillance data to allow CVO's of EU Member States to make rapid decisions in case of a future outbreak, as it would provide them with information on which areas and which time periods are likely to be vulnerable, which are likely to be safe and where the resources for surveillance should be allocated.

The RF models for the Pulicaris ensemble had poorer predictive power compared to Obsoletus ensemble and *C. imicola*. The abundance of the Pulicaris ensemble was ten-fold less than the abundance of the Obsoletus ensemble [33]. This lead to a lower number of Presence farms and, therefore, when the data were split into training and test sets, only a few Presence points were present in the test set. This resulted into heavily imbalanced monthly datasets e.g. February only included three farms with Presence observations in the test set. It is not recommended to assess model performance based only on a couple of observations from a certain class because it might lead to results with high variability. *Culicoides pulicaris* (*sensu stricto*) has been implicated in BTV transmission [67], but the Pulicaris ensemble species is not thought to have played a significant role in the 2006 BT outbreak in northern Europe [16]. Nevertheless, species of this ensemble might play a role in future outbreaks of emerging infections.

The model performance for *C. imicola* was highly accurate, with high AUC values for all months, indicating that this species has particular environmental requirements that can be detected through satellite imagery. This is likely to be related to hot and dry summers with low seasonal variation [64]: characteristic of the

Mediterranean basin. The three classes were clearly distinguishable in the maps, and Presence and Uncertain areas could be delimited to the Iberian Peninsula. *Culicoides imicola* maps can be used directly to allocate resources for surveillance programs or to determine appropriate animal movement restrictions.

In our models, the most important explanatory variables for classification of areas for the Presence/Absence affecting *C. imicola* distribution were related to temperature and precipitation. We found that during winter, the mean temperature of the coldest quarter was the variable driving the presence of *C. imicola*, while variables related to precipitation were the most predominant drivers during the warmer months. This is in accordance with the results of previous studies [56, 64, 68].

The distribution of *C. imicola* has previously been modelled at continental level using classical statistical models fitted to data collected from single European countries [57, 64, 69]. In our maps, *C. imicola* appeared to be present all year round, as it can be found on the southern coast of Spain during January and February. This agrees with previous analyses of the start of the vector season in Europe, where *C. imicola* was found to be present during the winter months in southern Spain and central and southern Portugal [65]. The predicted probability of presence shown in our maps are in agreement to the distribution models made for Spain by Ducheyne et al. [56], Calvete et al. [55] and Peters et al. [70], and for France, where the Presence areas for the species are mainly located in the coastal regions of Corsica and VAR department [15].

In our study, we used *Culicoides* data aggregated into groups, namely the Obsoletus and Pulicaris ensembles. Aggregating species into a single group, or ensemble, might represent a challenge for ecological modelling, as the different species might require different environmental conditions and phenology differ between them. This has been studied by Searle et al. [24], who estimated the start and end date of the vector season and length of the vector-free period for four species of the Obsoletus ensemble. They observed that there were differences in phenology among the species. The lower model performance obtained for Obsoletus and Pulicaris ensembles compared to *C. imicola* may reflect that different species within each ensemble have different

phenology and different environmental drivers. It would therefore be useful to identify *Culicoides* specimens to the species level. Molecular techniques, such as high-throughput real-time RT-PCR assays, can be used in a fast way for species identification. More accurate results could be expected if modelling is carried out on individual species data.

In practice, maps based on the classifications made for each 1 km^2 pixel might be difficult to use for decision making, as it becomes challenging to define classes for larger areas in which pixels from different classes are found. For practical use, predicted pixel values may therefore be summarized by area, such as at NUTS level (nomenclature of territorial units for statistics) defined by Eurostat (2013). This would facilitate the implementation of control and surveillance programs by European veterinary authorities.

Random Forest is a machine learning technique that has previously been used for ecological species modelling [19, 56, 70–75]. This technique has been proven to perform better compared to other applications of classical statistical methods such as Non-Linear Discriminant Analysis and Generalized Ginear Models [19, 71], as well as Linear Discriminant Analysis, logistic regression [70, 74] and Additive Logistic Regression [75]. In this work, the monthly predicted probability of *Culicoides* presence had medium-high accuracy, but it is important to keep in mind that there might be other variables that cannot be captured by satellite imagery and that may have an influence on the occurrence of these species on a local scale, such as soil conditions (affecting breeding sites) and farming practices. Nevertheless, for some months, our models performed slightly better than other RF models used for predicting the occurrence of biting midges and mosquitoes [70, 71]. This highlights the challenges faced in predicting the occurrence of insect vectors using remote sensing data, as vectors are highly influenced by local microenvironments [76] and these data are difficult to obtain from satellite images without high spatial resolution.

Conclusions

We present here maps as a risk assessment tool that can be used in the future to predict potential risk areas and risk seasons for *Culicoides*-borne disease outbreaks. They are particularly useful for European veterinary authorities, who can classify both areas likely to have vectors and likely to be vector-free in advance and during a sudden outbreak in areas without active entomological surveillance. Predicting areas of uncertain status allows focusing costly active entomological surveillance to limited areas. The developed gain functions used to delimit the areas for targeted active surveillance can easily be adjusted to new diseases where the cost of concluding false presence or false absence may be different than suggested here.

Additional files

Additional file 1: Table S1. The top five of the most important variables by species group for each month. The variable importance is scaled from 0 to 100. Within each month (columns), the most important variable has a value of 100.

Abbreviations

BT: Bluetongue disease; BTV: Bluetongue virus; SVFP: Seasonal vector-free period; RF: Random Forest; MIR: Mid-infrared, dLST: Daytime land surface temperature; nLST: Nighttime land surface temperature; NDVI: Normalized difference vegetation index; TFA: Temporal Fourier analysis; CLC: Corine Land Cover; PP: Probability of presence; CV: Cross-validation, AUC: Area under the ROC curve; ROC: Receiver operating characteristics curve; CVO: Chief Veterinary Officer ; NUTS: Nomenclature of territorial units for statistics; Pres: Presence class; Abs: Absence class

Acknowledgments

We would like to thank the Direction Générale de l'Alimentation from the French Ministry in charge of agriculture for the funding, and the Directions départementales de la protection des populations for their support in collecting the biting midges during the survey. We also thank the Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente for providing data on the national surveillance of *Culicoides* in Spain.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The national surveillance and research data that support the findings of this study are available from the following people: Spain, Miguel Ángel Miranda Chueca; France, Thomas Balenghien; Germany, Jörn Gethmann; Denmark, Rene Bødker; Sweden, Anders Lindström; Norway, Petter Hopp; Poland, Magdalena Larska; Austria, Katharina Brugger; Switzerland, Alexander Mathis. Restrictions apply to the availability of these data, which were used under license for the current study and are not publicly available. Data are, however, available from the authors upon reasonable request and with permission from the national surveillance programmes of each country.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the EMIDA ERA-NET-supported project VICE (Vector-borne Infections: Risk-based and Cost-Effective Surveillance Systems). *Culicoides* data from

Germany were partly collected within the German part of the VICE project funded by EMIDA ERA-NET through the Federal Office for Agriculture and Food (grant no. 314-06.01-2811ERA248). The Swiss Food Safety and Veterinary Office and the Vet-Austria project provided financial support to the Swiss and Austrian partners, respectively.

Authors' contributions

ACC analysed the data and drafted the manuscript. RB planned the original study and contributed to the analysis and drafting the manuscript. LJK contributed to the analysis and drafting of the manuscript. AB and AS contributed to the analysis and interpretation of the results. HS, SAN, MGA, AL, JC, RL, SS, EK, JG, FC, ML, MS, AO, IH, SS, PH, KB, FR, TB, CG, IR, XA, JL, JCD, BM, DD, MLS, RV, BS, MAMC, CB, JL, RE, AM and WT discussed and identified preliminary national data, negotiated data access with national surveillance authorities and research projects, jointly discussed the taxonomic and spatial resolution for analysis and identified, selected and extracted the final data and relevant variables that met the criteria for the joint database. They also wrote the protocol summaries and commented on the analysis results and edited the manuscript. All authors read and approved the final manuscript.

Author details

[1]Division for Diagnostics and Scientific Advice, National Veterinary Institute, Technical University of Denmark (DTU), Lyngby, Denmark. [2]Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Lyngby, Denmark. [3]Department of Agroecology - Entomology and Plant Pathology, Aarhus University, Aarhus, Denmark. [4]Department of Science and Environment, Roskilde University, Roskilde, Denmark. [5]National Veterinary Institute (SVA), Uppsala, Sweden. [6]Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Reference and Research National Reference Centre for Tropical Infectious Diseases, Hamburg, Germany. [7]Department of Biology and Environmental Sciences, Carl von Ossietzky University, Oldenburg, Germany. [8]Institute of

Epidemiology, Friedrich Loeffler Institute, Greifswald, Germany. [9]Department of Virology, National Veterinary Research Institute, Pulawy, Poland. [10]Norwegian Veterinary Institute, Oslo, Norway. [11]Institute for Veterinary Public Health, Vetmeduni, Vienna, Austria. [12]CIRAD, UMR ASTRE, F-34398 Montpellier, France. [13]Institute of Parasitology and Tropical Pathology of Strasbourg, EA7292, Université de Strasbourg, Strasbourg, France. [14]EID Méditerranée, Montpellier, France. [15]Laboratory of Zoology, University of the Balearic Islands, Palma, Spain. [16]Department of Animal Pathology, University of Zaragoza, Zaragoza, Spain. [17]Institute of Parasitology, University of Zürich, Zürich, Switzerland. [18]Avia-GIS NV, Zoersel, Belgium.

References

1. Du Toit RM. The transmission of blue-tongue and horse-sickness by *Culicoides*. Onderstepoort J Vet Sci Anim Ind. 1944;19:7–16.

2. Elbers ARW, Meiswinkel R, van Weezep E, Sloet van Oldruitenborgh-Oosterbaan MM, Kooi EA. Schmallenberg virus in *Culicoides* spp. biting midges, the Netherlands, 2011. Emerg Infect Dis. 2013;19:106–9.

3. Mellor PS, Boned J, Hamblin C, Graham S. Isolations of African horse sickness virus from vector insects made during the 1988 epizootic in Spain. Epidemiol Infect. 1990;105:447–54.

4. Mellor PS, Carpenter S, Harrup L, Baylis M, Mertens PPC. Bluetongue in Europe and the Mediterranean Basin: History of occurrence prior to 2006. Prev Vet Med. 2008;87:4–20.

5. Carpenter S, Wilson A, Mellor PS. *Culicoides* and the emergence of bluetongue virus in northern Europe. Trends Microbiol. 2009;17:172–8.

6. Toussaint J-F, Sailleau C, Mast J, Houdart P, Czaplicki G, Demeestere L, et al. Bluetongue in Belgium, 2006. Emerg Infect Dis. 2007;13:614–6.

7. Thiry E, Saegerman C, Guyot H, Kirten P, Losson B, Rollin F, et al. Bluetongue in northern Europe. Vet Rec. 2006;159:327–7.

8. Mehlhorn H, Walldorf V, Klimpel S, Jahn B, Jaeger F, Eschweiler J, et al. First occurrence of *Culicoides obsoletus*-transmitted bluetongue virus epidemic in central Europe. Parasitol Res. 2007;101:219–28.
9. Zientara S, Sánchez-Vizcaíno JM. Control of bluetongue in Europe. Vet Microbiol. 2013;165:33–7.
10. Pinior B, Brugger K, Kofer J, Schwermer H, Stockreiter S, Loitsch A, et al. Economic comparison of the monitoring programmes for bluetongue vectors in Austria and Switzerland. Vet Rec. 2015;176:464–464.
11. Rushton J, Lyons N. Economic impact of bluetongue: a review of the effects on production. Vet Ital. 2015;51:401–6.
12. Hoffmann B, Bauer B, Bauer C, Bätza HJ, Beer M, Clausen PH, et al. Monitoring of putative vectors of bluetongue virus serotype 8, Germany. Emerg Infect Dis. 2009;15:1481–4.
13. Carpenter S, Mcarthur C, Selby R, Ward R, Nolan DV, Mordue Luntz AJ, et al. Experimental infection studies of UK *Culicoides* species midges with bluetongue virus serotypes 8 and 9. Vet Rec. 2008;163:589–92.
14. Dijkstra E, van der Ven IJK, Meiswinkel R, Holzel DR, van Rijn PA, Meiswinkel R. *Culicoides chiopterus* as a potential vector of bluetongue virus in Europe. Vet Rec. 2008;162:422–422.
15. Venail R, Balenghien T, Guis H, Tran A, Setier-Rio M-L, Delécolle J-C, et al. Assessing diversity and abundance of vector populations at a national scale: example of *Culicoides* surveillance in France after bluetongue virus emergence. In: Mehlhorn H, editor. Arthropods as Vectors. Arthropods as Vectors of Emerging Diseases. Berlin-Heidelberg: Springer; 2012. p. 77–102.
16. Meiswinkel R, Baldet T, de Deken R, Takken W, Delécolle J-C, Mellor PS. The 2006 outbreak of bluetongue in northern Europe - the entomological perspective. Prev Vet Med. 2008;87:55–63.
17. Hartemink N, Vanwambeke SO, Purse B V, Gilbert M, Van Dyck H. Towards a resource-based habitat approach for spatial modelling of vector-borne disease risks. Biol

Rev. 2015;90:1151–62.
18. Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. Global environmental data for mapping infectious disease distribution. Adv Parasitol. 2006;62:37–77.
19. Cianci D, Hartemink N, Ibáñez-Justicia A. Modelling the potential spatial distribution of mosquito species using three different techniques. Int J Health Geogr. 2015;14:10.
20. Kalluri S, Gilruth P, Rogers D, Szczur M. Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. PLoS Pathog. 2007;3:e116.
21. EFSA Panel on Animal Health and Welfare. Bluetongue: control, surveillance and safe movement of animals. EFSA J. 2017;15: 3.
22. European Commission. Ec 1266/2007. Off J Eur Union. 2007; L 283: 37-52.

23. Brugger K, Köfer J, Rubel F. Outdoor and indoor monitoring of livestock-associated *Culicoides* spp. to assess vector-free periods and disease risks. BMC Vet Res. 2016;12:88.
24. Searle KR, Barber J, Stubbins F, Labuschagne K, Carpenter S, Butler A, et al. Environmental drivers of *Culicoides* phenology: how important is species-specific variation when determining disease policy? PLoS One. 2014;9:e111876.
25. Kaufmann C, Steinmann IC, Hegglin D, Schaffner F, Mathis A. Spatio-temporal occurrence of *Culicoides* biting midges in the climatic regions of Switzerland, along with large scale species identification by MALDI-TOF mass spectrometry. Parasit Vectors. 2012;5:246.
26. Ander M, Meiswinkel R, Chirico J. Seasonal dynamics of biting midges (Diptera: Ceratopogonidae: *Culicoides*), the potential vectors of bluetongue virus, in Sweden. Vet Parasitol. 2012;184:59–67.
27. Mehlhorn H, Walldorf V, Klimpel S, Schmahl G, Al-Quraishy S, Walldorf U, et al. Entomological survey on vectors of nluetongue virus in Northrhine-Westfalia (Germany) during 2007 and 2008. Parasitol Res. 2009;105:321–9.

28. Clausen P-H, Stephan A, Bartsch S, Jandowsky A, Hoffmann-Köhler P, Schein E, et al. Seasonal dynamics of biting midges (Diptera: Ceratopogonidae, *Culicoides* spp.) on dairy farms of central Germany during the 2007/2008 epidemic of bluetongue. Parasitol Res. 2009;105:381–6.
29. Kiel E, Liebisch G, Focke R, Liebisch A. Monitoring of *Culicoides* at 20 locations in northwest Germany. Parasitol Res. 2009;105:351–7.
30. Afonso A, Abrahantes JC, Conraths F, Veldhuis A, Elbers A, Roberts H, et al. The Schmallenberg virus epidemic in Europe - 2011–2013. Prev Vet Med. 2014;116:391–403.
31. Hoffmann B, Scheuch M, Höper D, Jungblut R, Holsteg M, Schirrmeier H, et al. Epizootic of ovine congenital malformations associated with Schmallenberg virus infection. Emerg Infect Dis. 2012;18:469–72
32. Ortega MD, Mellor PS, Rawlings P, Pro MJ. The seasonal and geographical distribution of *Culicoides imicola*, *C. pulicaris* group and *C. obsoletus* group biting midges in central and southern Spain. Arch Virol Suppl. 1998;14:85-91.
33. Cuéllar AC, Kjær LJ, Kirkeby C, Skovgard H, Nielsen SA, Stockmarr A, et al. Spatial and temporal variation in the abundance of *Culicoides* biting midges (Diptera: Ceratopogonidae) in nine European countries. Parasit Vectors. 2018;11:112.
34. Nielsen SA, Nielsen BO, Chirico J. Monitoring of biting midges (Diptera: Ceratopogonidae: *Culicoides* Latreille) on farms in Sweden during the emergence of the 2008 epidemic of bluetongue. Parasitol Res. 2010;106:1197–203.
35. Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, Wint GRW, et al. Global data for ecology and epidemiology: a novel algorithm for temporal fourier processing MODIS data. PLoS One. 2008;3:e1408.
36. EDENext. Biology and control of vector-borne infections in Europe. 2011. https://www.edenext.eu/. Accessed 28 Oct2018.
37. Hijmans RJ. Worldclim - Global Climate Data. Free climate data for ecological modeling and GIS.2005. http://www.worldclim.org/node/1. Accessed 28 Oct 2018.
38. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution

interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965–78.
39. European Environment Agency. Corine Land Cover. 2018. https://www.eea.europa.eu/data-and-maps/data/clc-2006-raster-4. Accessed 28 Oct 2018.
40. Robinson TP, Wint GRW, Conchedda G, Van Boeckel TP, Ercoli V, Palamara E, et al. Mapping the global distribution of livestock. PLoS One. 2014;9:5.
41. Breiman L. Random Forests. Mach. Learn. 2001;45:5–32.
42. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer; 2013.
43. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
44. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28:159–60.
45. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2/3:18–22.
46. Breiman L. Statistical modeling: The two cultures. Stat Sci. 2001;16:199–231.
47. Guis H, Caminade C, Calvete C, Morse AP, Tran A, Baylis M. Modelling the effects of past and future climate on the risk of bluetongue emergence in Europe. J R Soc Interface 2012;9:339–50.
48. Pearce J, Ferrier S. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol Modell. 2000;133:225–45.
49. Lunardon N, Menardi G, Torelli N. ROSE : A package for binary imbalanced learning. R J. 2014;6:79–89.
50. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6:20–9.
51. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27:861–74.
52. Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecol Modell. 2000;135:147–86.
53. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010;5:1315–6.

54. Liu C, Berry PM, Dawson TP, Pearson RG. Selecting thresholds of occurrence in the prediction of species distributions. Ecography. 2005;28:385–93.
55. Calvete C, Estrada R, Miranda MA, Borrás D, Calvo JH, Lucientes J. Modelling the distributions and spatial coincidence of bluetongue vectors *Culicoides imicola* and *the Culicoides obsoletus* group throughout the Iberian peninsula. Med Vet Entomol. 2008;22:124–34.
56. Ducheyne E, Miranda Chueca MA, Lucientes J, Calvete C, Estrada R, Boender G, et al. Abundance modelling of invasive and indigenous *Culicoides* species in Spain. Geospat Health 2013;8:241–54.
57. Purse BV, Tatem AJ, Caracappa S, Rogers DJ, Mellor PS, Baylis M, et al. Modelling the distributions of *Culicoides* bluetongue virus vectors in Sicily in relation to satellite-derived climate variables. Med Vet Entomol. 2004;18:90–101.
58. Purse BV, Brown HE, Harrup L, Mertens PPC, Rogers DJ. Invasion of bluetongue and other orbivirus infections into Europe: the role of biological and climatic processes. Rev Sci Tech. 2008;27:427–42.
59. Brugger K, Rubel F. Characterizing the species composition of European *Culicoides* vectors by means of the Köppen-Geiger climate classification. Parasit Vectors. 2013;6:333.
60. Purse BV, Carpenter S, Venter GJ, Bellis G, Mullens BA. Bionomics of temperate and tropical *Culicoides* midges: knowledge gaps and consequences for transmission of *Culicoides*-borne viruses. Annu Rev Entomol. 2015;60:373–92.
61. Lühken R, Steinke S, Hoppe N, Kiel E. Effects of temperature and photoperiod on the development of overwintering immature *Culicoides chiopterus* and *C. dewulfi*. Vet Parasitol. 2015;214:195–9.
62. Steinke S, Lühken R, Balczun C, Kiel E. Emergence of *Culicoides obsoletus* group species from farm-associated habitats in Germany. Med Vet Entomol. 2016;30:174–84.
63. EFSA. A story map. Bluetongue virus (BTV). 2017. https://efsa.maps.arcgis.com/apps/MapJournal/index.html?appid=80efdcdeb24646ccaa9bd28c7a343b42#. Accessed 28 Oct 2018.

64. Purse B, Mccormick BJJ, Mellor PS, Baylis M, Boorman JPT, Borras D, et al. Incriminating bluetongue virus vectors with climate envelope models. J Appl Ecol. 2007;44:1231–42.

65. Ramilo DW, Nunes T, Madeira S, Boinas F, da Fonseca IP. Geographical distribution of *Culicoides* (Diptera: Ceratopogonidae) in mainland Portugal: presence/absence modelling of vector and potential vector species. PLoS One. 2017;12:e0180606.

66. Calvete C, Estrada R, Miranda MA, Borrás D, Calvo JH, Lucientes J. Ecological correlates of bluetongue virus in Spain: predicted spatial occurrence and its relationship with the observed abundance of the potential *Culicoides* spp. vector. Vet J. 2009;182:235–43.

67. Caracappa S, Torina A, Guercio A, Vitale F, Calabrò A, Purpari G, et al. Identification of a novel bluetongue virus vector species of *Culicoides* in Sicily. Vet Rec. 2003;153:71–4.

68. Wittmann EJ, Mellor PS, Baylis M. Using climate data to map the potential distribution of *Culicoides imicola* (Diptera: Ceratopogonidae) in Europe. Rev Sci Tech. 2001;20:731–40.

69. Tatem AJ, Baylis M, Mellor PS, Purse BV, Capela R, Pena I, et al. Prediction of bluetongue vector distribution in Europe and north Africa using satellite imagery. Vet Microbiol. 2003;97:13–29.

70. Peters J, De Baets B, Van Doninck J, Calvete C, Lucientes J, De Clercq EM, et al. Absence reduction in entomological surveillance data to improve niche-based distribution models for *Culicoides imicola*. Prev Vet Med. 2011;100:15–28.

71. Ibañez-Justicia A, Cianci D. Modelling the spatial distribution of the nuisance mosquito species *Anopheles plumbeus* (Diptera: Culicidae) in the Netherlands. Parasit Vectors. 2015;8:258.

72. Ducheyne E, Charlier J, Vercruysse J, Rinaldi L, Biggeri A, Demeler J, et al. Modelling the spatial distribution of *Fasciola hepatica* in dairy cattle in Europe. Geospat Health. 2015;9:261–70.

73. Selemetas N, Ducheyne E, Phelan P, O'Kiely P, Hendrickx G, de Waal T. Spatial analysis and risk mapping of *Fasciola hepatica* infection in dairy herds in Ireland. Geospat Health. 2015;9:281–91.

74. van Doninck J, De Baets B, Peters J, Hendrickx G, Ducheyne E, Verhoest NEC. Modelling the spatial distribution of *Culicoides imicola*: Climatic versus remote sensing data. Remote Sens. 2014;6:6604–19.

75. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random Forests for classification in ecology. Ecology. 2007;88:2783–92.

76. Haider N, Cuellar AC, Kjær LJ, Sørensen JH, Bødker R. Microclimatic temperatures at Danish cattle farms, 2000–2016: Quantifying the temporal and spatial variation in the transmission potential of Schmallenberg virus. Parasit Vectors. 2018;11:128.

3.2.1 Unpublished results relating Manuscript II

In this section, the results of analysing different balancing methods are presented. The aim of this section is to clarify the reason for using Up scaling as balancing method in the second manuscript (chapter 3: Results).

I show the results obtained from different balancing methods and respond to three questions that came up during the analysis:

1. Is the AUC sensitive to random data partitions (comparison between data partitions)?

2. Which method has the best performance in terms of AUC score, considering the mean AUC and also the magnitude of the variation of the predicted AUC for different seeds and months?

3. Are the results obtained using the best balancing method significantly different from the ones using the imbalanced training set?

Class imbalance and balancing methods

The monthly *Culicoides* datasets used for this thesis were highly imbalanced, meaning that they had a larger proportion of one class (presence or absence) compared to the other class. Class imbalance represents a problem for machine learning techniques as the techniques may be very good at predicting the majority class (the class with the larger proportion) but less so when predicting the minority class (the class with the smaller proportion). A classifier that optimizes by accuracy, such as classification trees, shows a high accuracy despite the minority class being misclassified [Chawla et al., 2002, Japkowicz and Stephen, 2002, Kuhn and Johnson, 2013, Prati et al., 2004, Weiss and Provost, 2001]. In the monthly datasets, the proportion of presences was relatively low compared to absences, and during the summer we found a higher proportion of presences compared to absences.

To overcome a class imbalance problem, balancing the class distribution to equal proportions has been proposed as an adequate solution to improve the performance of a classifier.Different resampling techniques are available for balancing class distributions in a dataset.

I only used the Obsoletus ensemble to assess model performance comparing the unbalanced training set against five different balancing methods (SMOTE [Chawla et al., 2002], ROSE [Lunardon et al., 2014], oversampling, down-sampling, tomek [Batista et al., 2004]).

For this thesis, I (i) analysed the effect of using the five balancing methods together with an unbalanced version of the training sets, all derived from the same data set and (ii) compared the effects of using different data partitions of that dataset within each individual method. This was done for each month as follows:

1. Each monthly dataset was divided into a training and test set. The data was split automatically using the createDataPartition of the caret package in R. This function performs a random sampling within each group of the observed classes ("Presence" and "Absence") keeping the same ratio between the classes within the training and test sets.

2. The training set was copied five times and each of five copies was balanced using the different balancing methods. The original training set was retained as an unbalanced version. The test set remained unbalanced as well. A random forest model (RF) was run for each of these six monthly training sets.

3. For each of the six RF models obtained this way, I predicted the probability of presence (PP) for the test set samples. As the test set remained unbalanced, it represents the class distribution of the original data.

4. I repeated this step 10 times, using 10 different random data partitions. As consequence of this, for each month, 10 different pairs of train and test sets were obtained (60 training sets and 10 test sets) per month. Before balancing, all the 70 monthly datasets contained the same ratio of the two classes (only small differences are produced by the random process involved in the data partition).

5. In total 720 RF models were run (12 months x (1 non-balance +5 balancing methods) x 10 data partitions).

The next figures show the results obtained. In each plot, the frequency of samples for each class in the test set (Y axis) was plotted as a function

of the probability of presence predicted by the models for those samples (X-axis). For practical reasons, here I show the results obtained for only six of the ten data partitions (seed 1 to seed 6). Different data partitions are in different columns while the different balancing methods are found in each row.

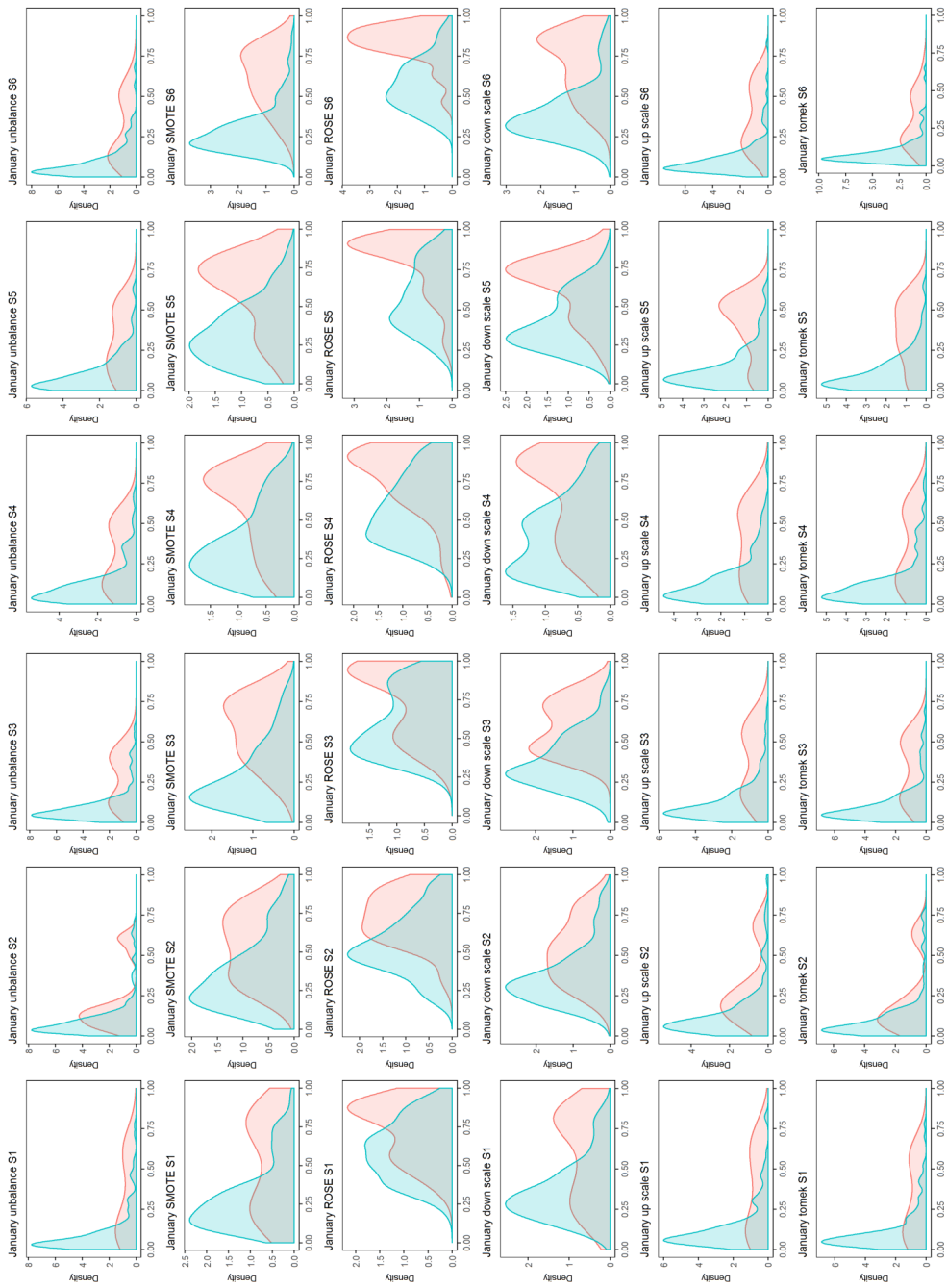


Figure 3.1: January, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

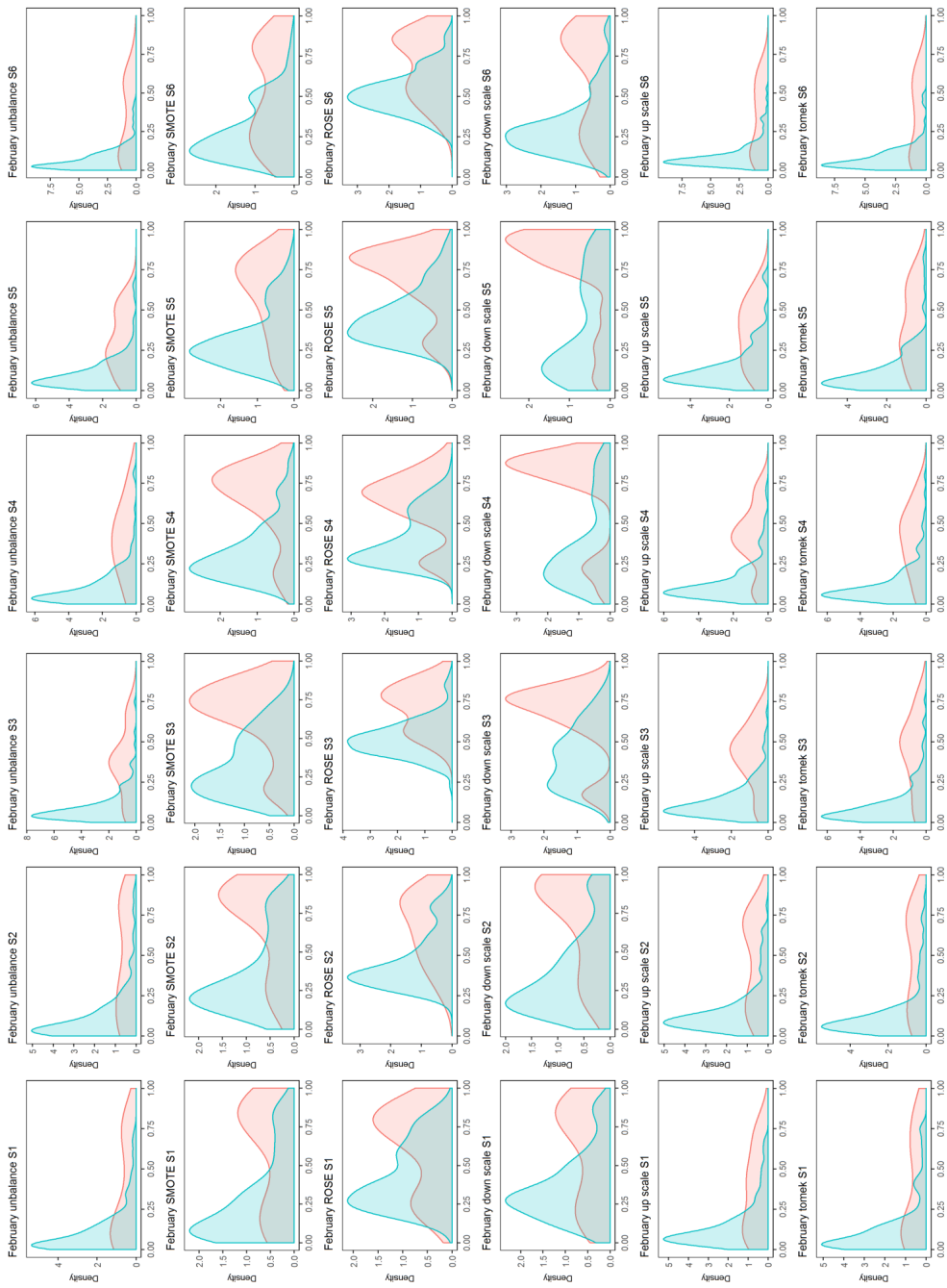


Figure 3.2: February, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

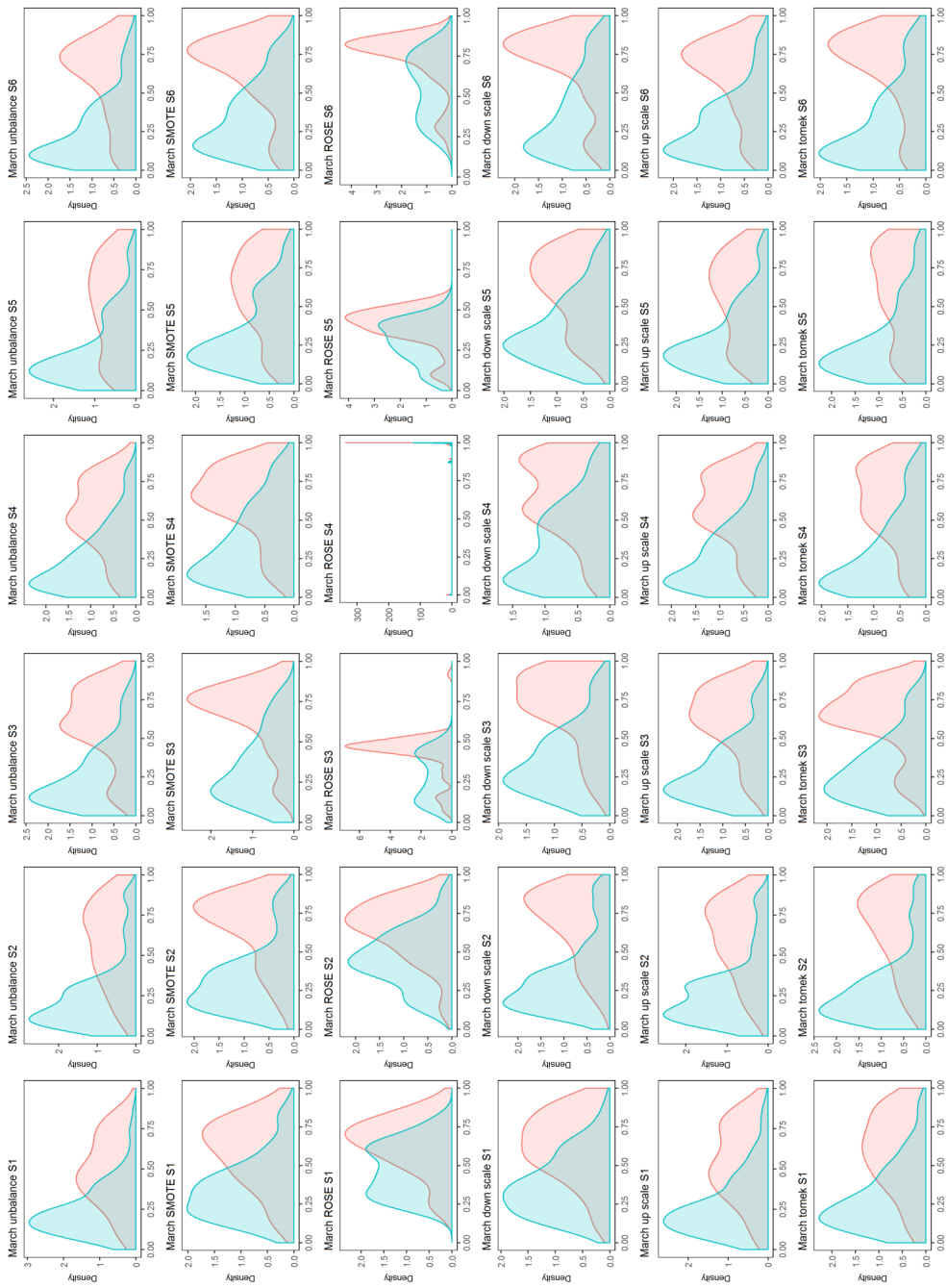


Figure 3.3: March, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

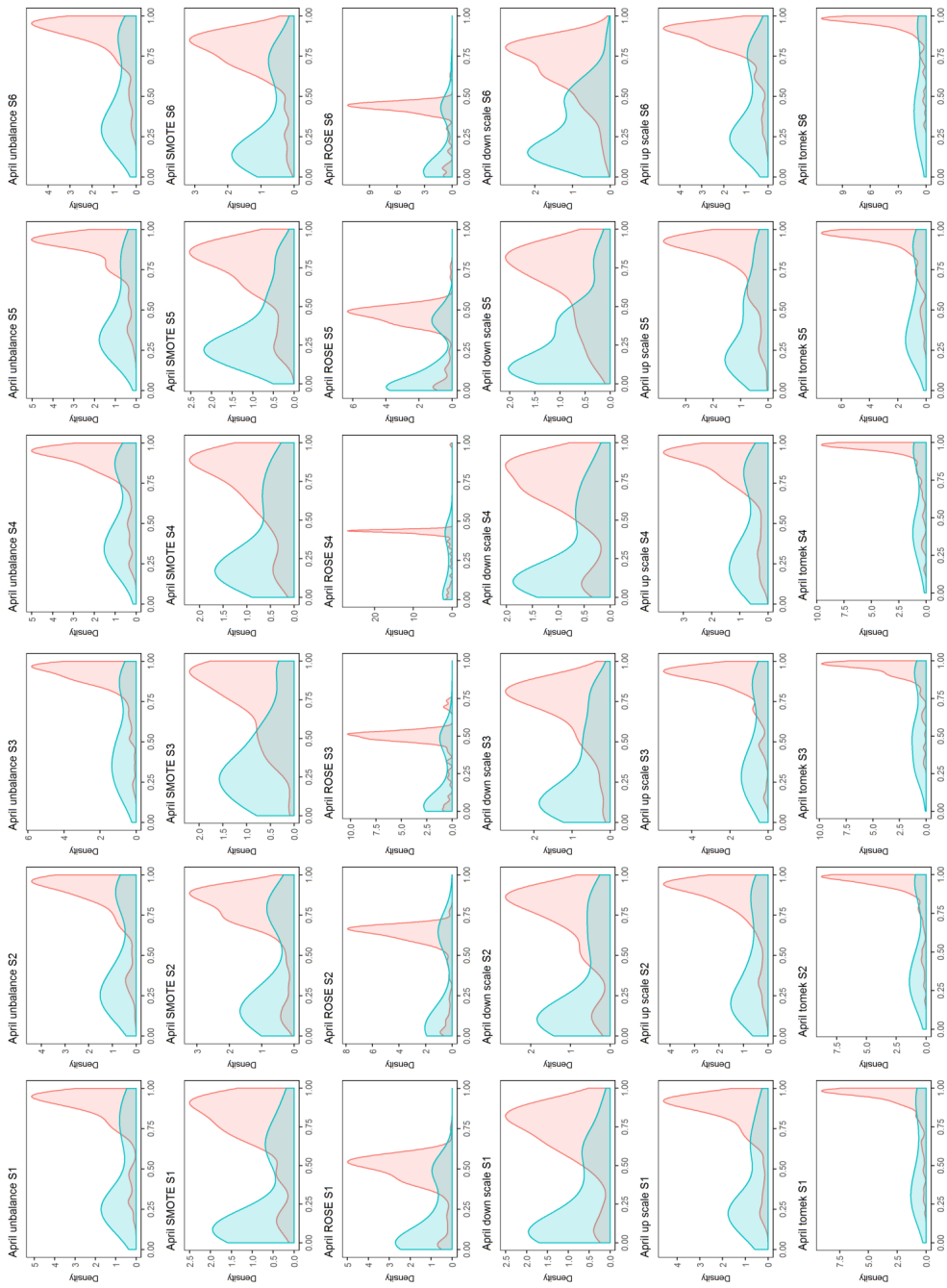


Figure 3.4: April, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

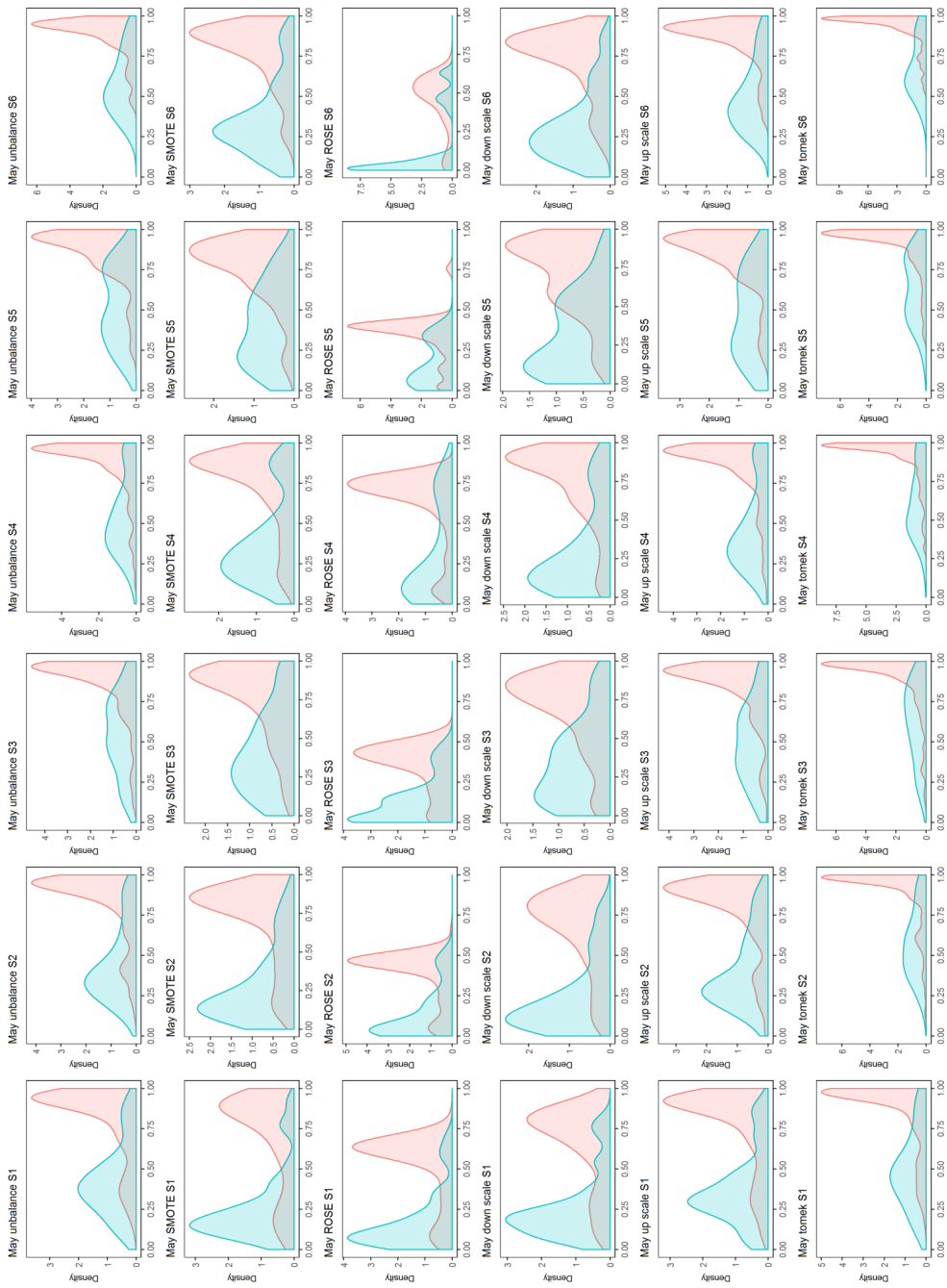


Figure 3.5: May, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

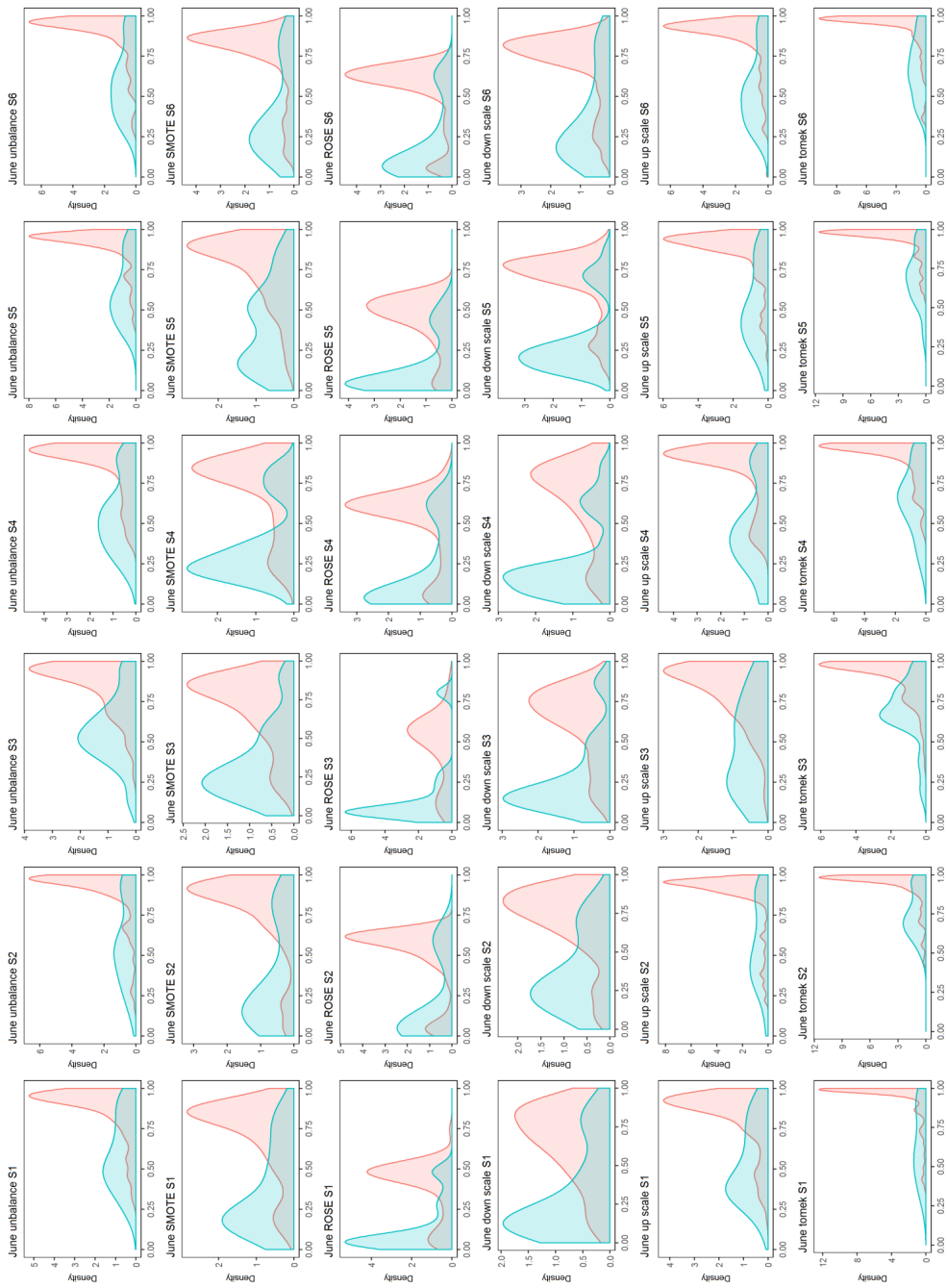


Figure 3.6: June, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

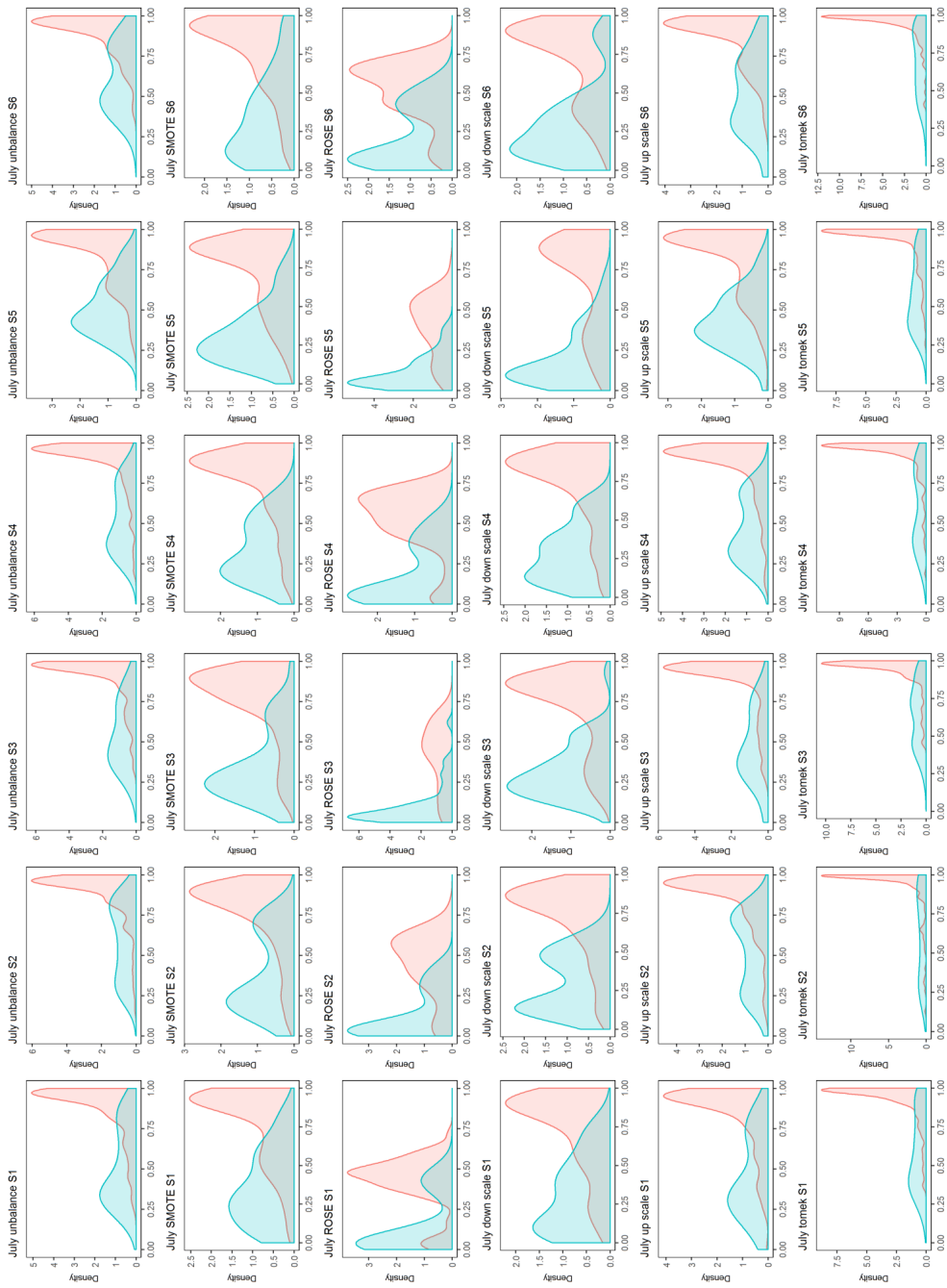


Figure 3.7: July, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

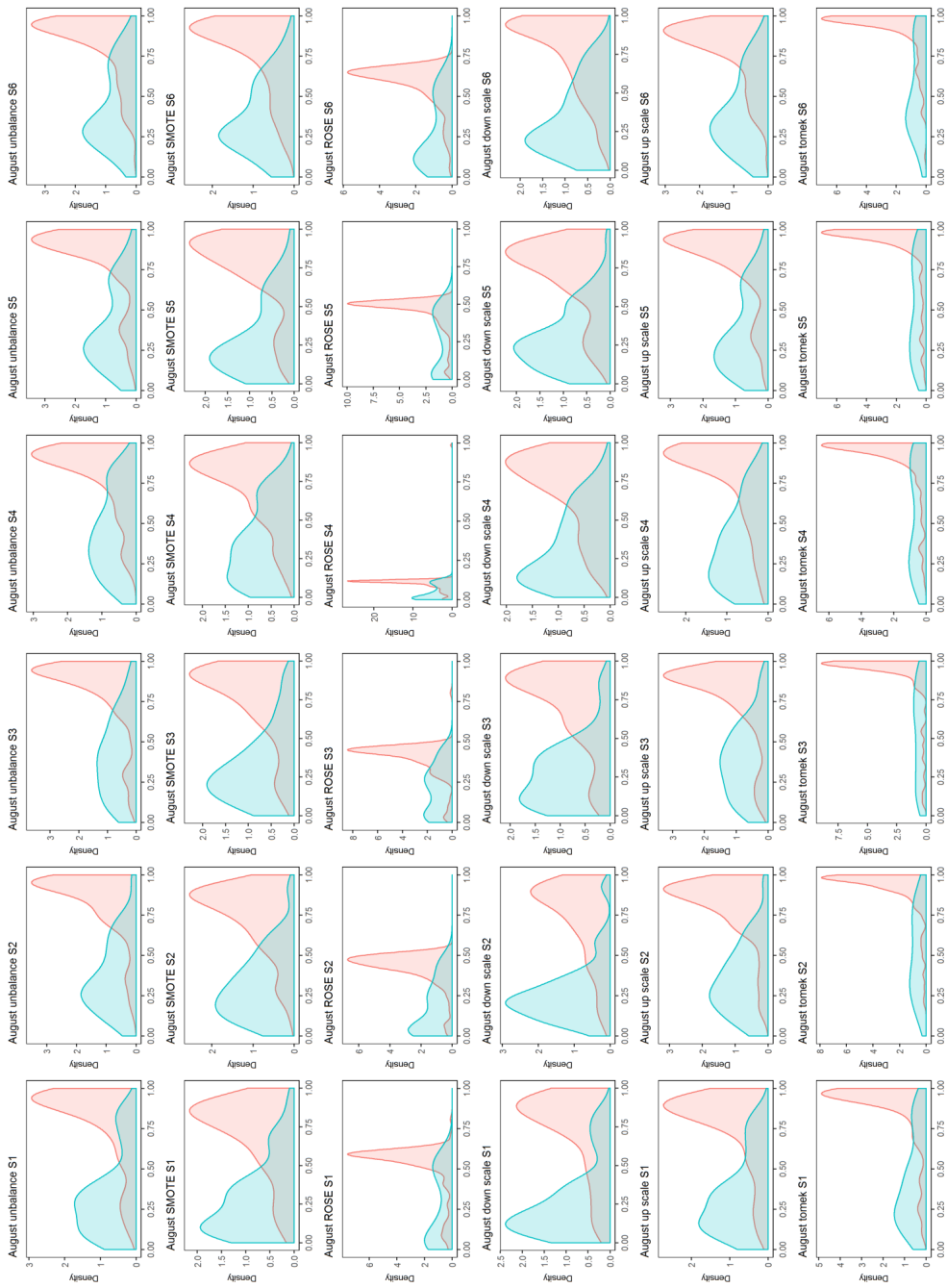


Figure 3.8: August, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

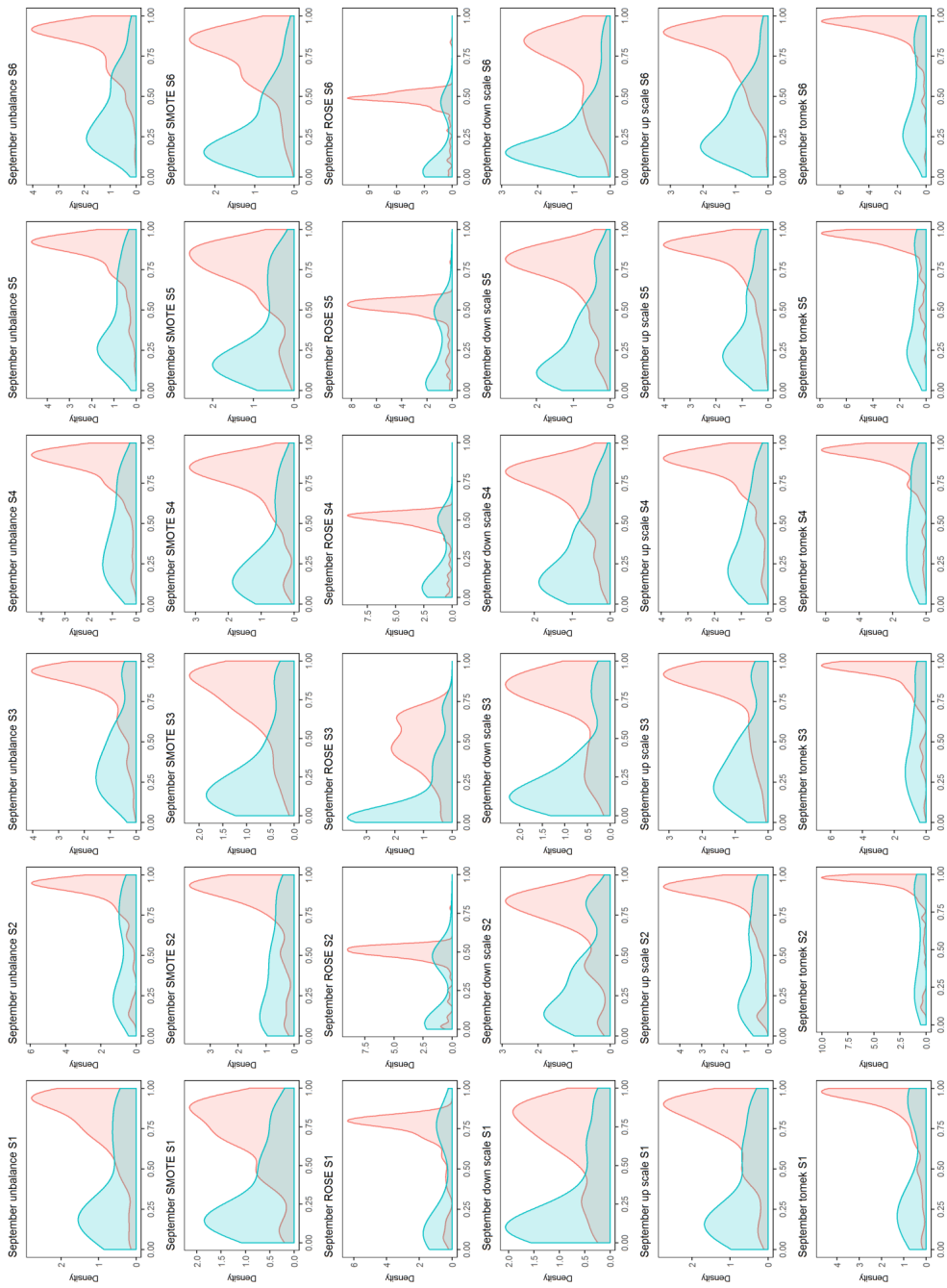


Figure 3.9: September, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

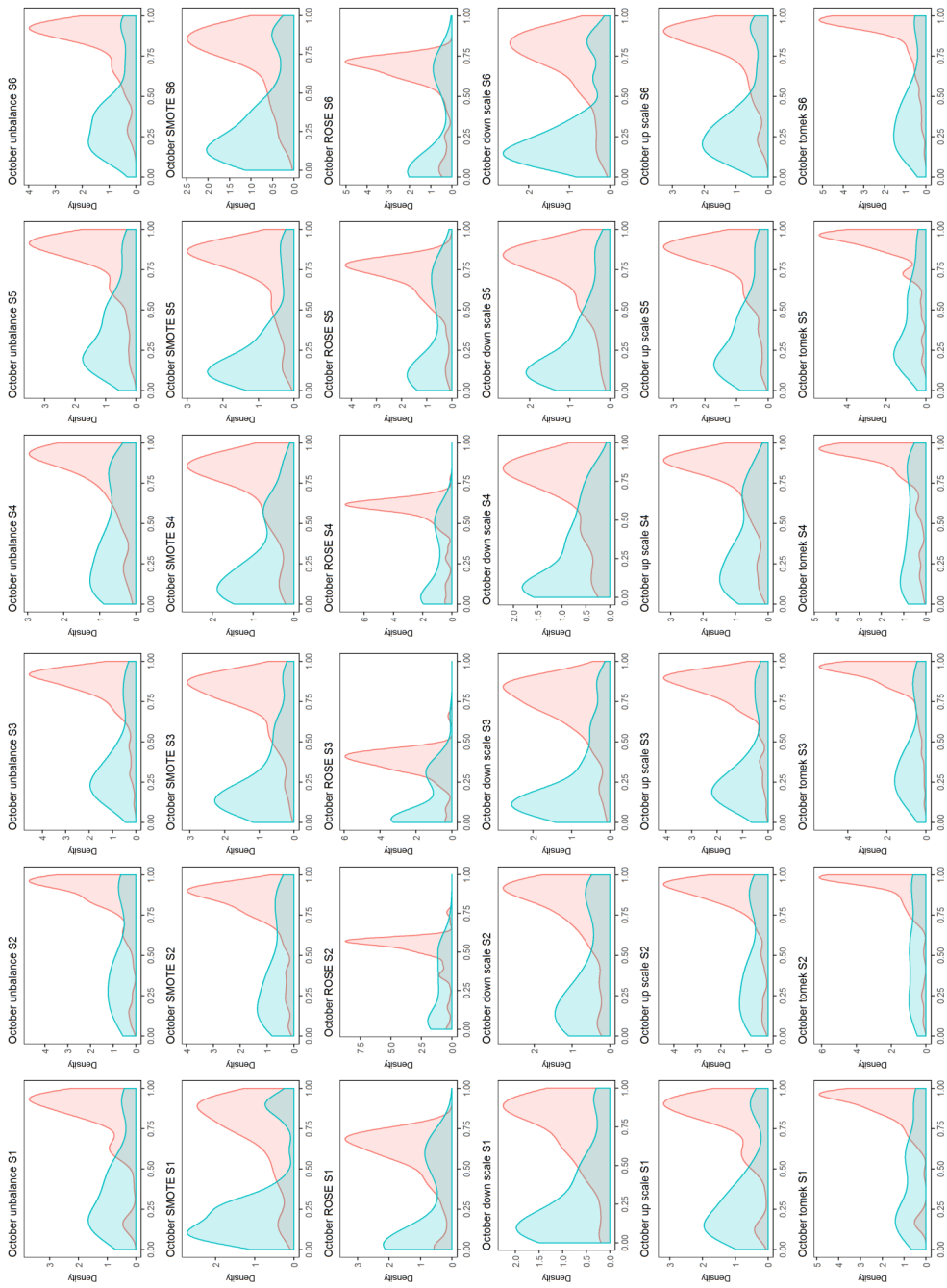


Figure 3.10: October, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

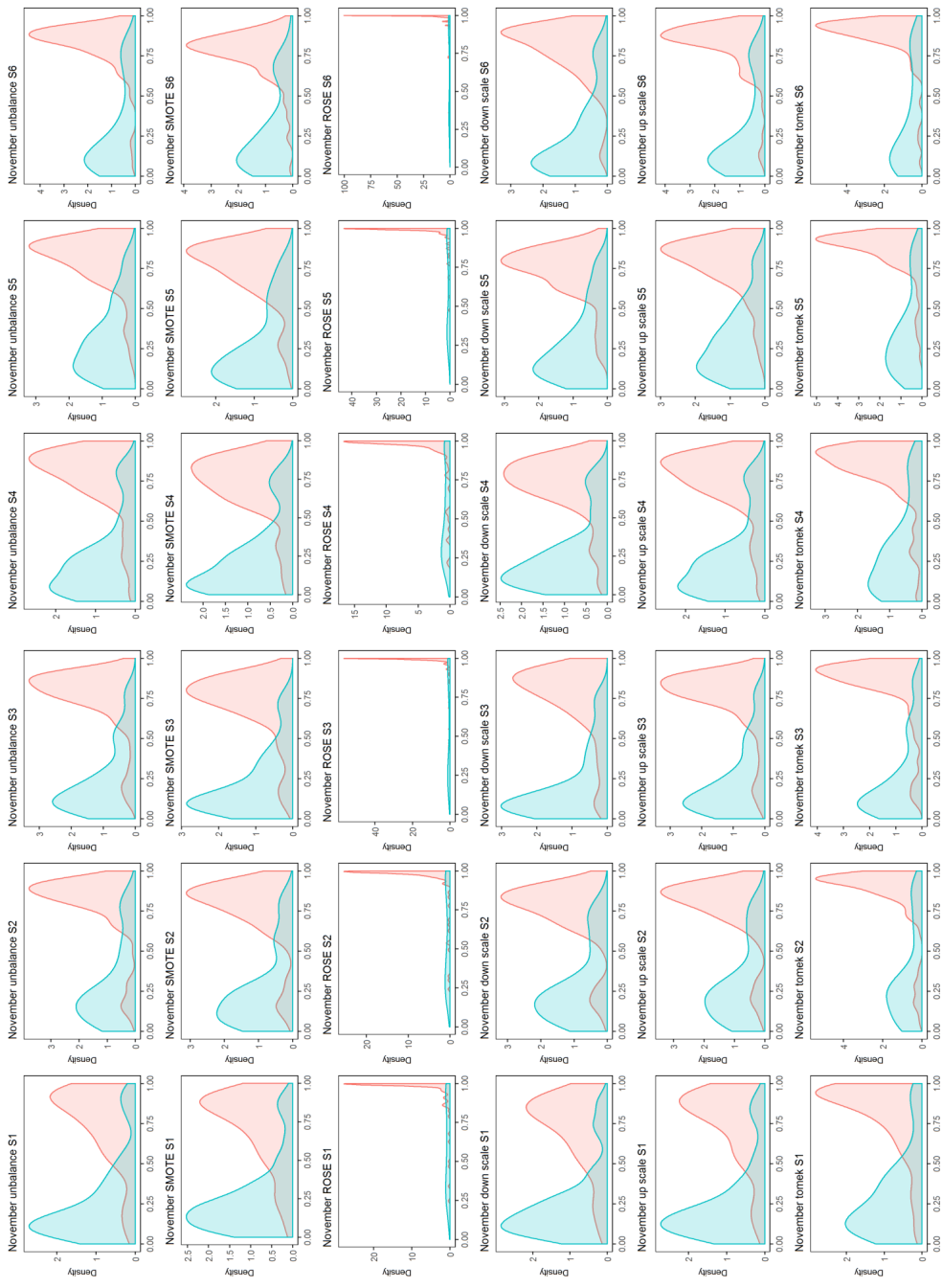


Figure 3.11: November, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

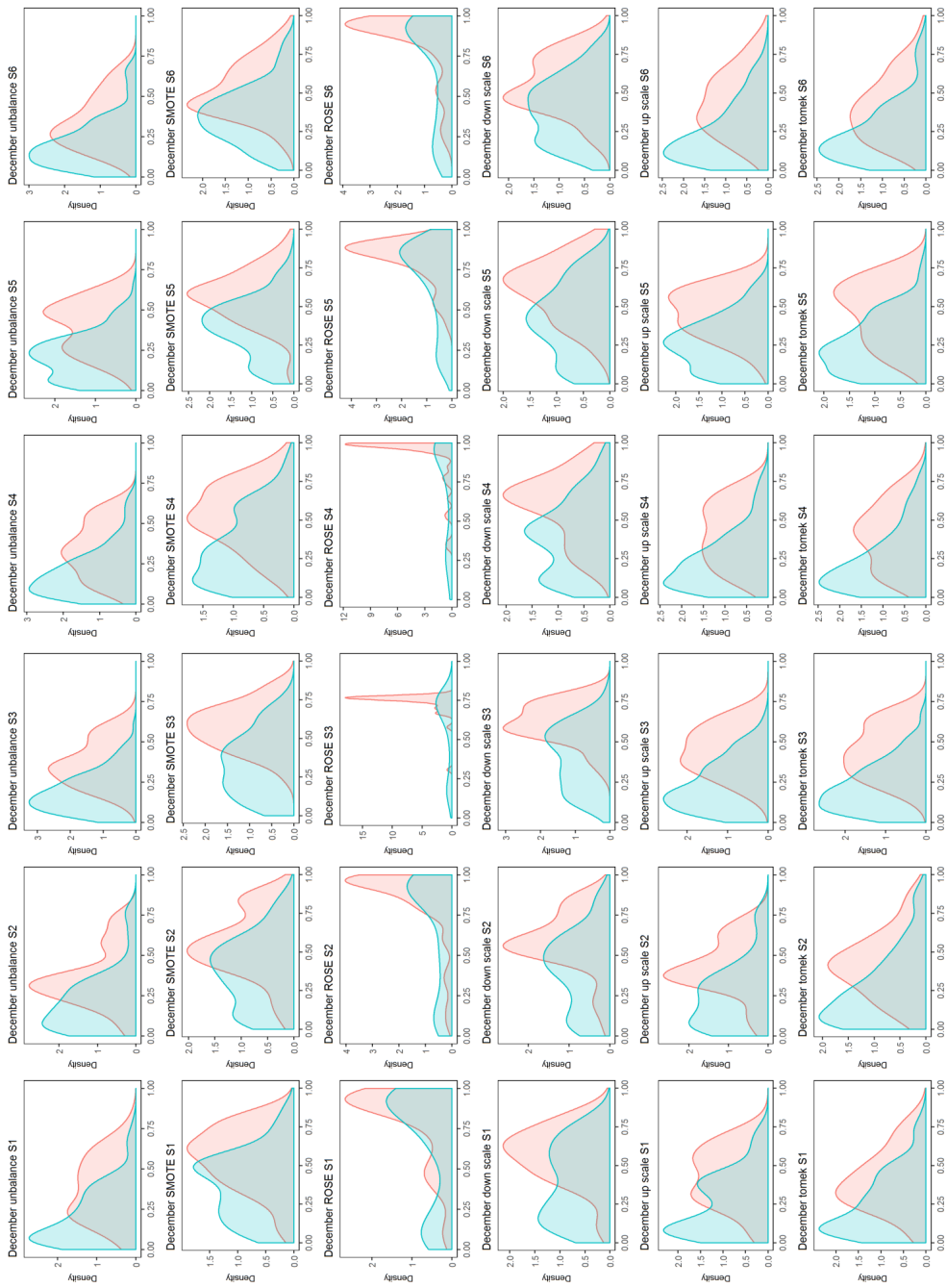


Figure 3.12: December, density function (Y - axis) as a function of the predicted probability of presence (X - axis). Blue: Absence class. Red: Presence class. This analysis was made for 10 seeds (data partitions) but only results for first six seeds are displayed (S1-S6).

In general, different data partitions produced very similar results for each month. This low variation applied to all of the balancing methods.

Question 1: Does the AUC differ among different data partitions (comparison among seeds within a single method)?

Method: For each of the 720 models obtained, the AUC value was computed by calculating the ROC curve for the test. Per month, I evaluated if the resulting AUC depended on the ten different random data partitions. The distribution of the ten obtained AUC values for each balancing method was then plotted for each month (Figure **??**).

Results: see Figure 3.13.

Result and conclusion: For some months, data partition had a certain effect in the model performance, as the different AUC values obtained from different seeds showed. For instance in January, there was a seed that outperformed compared to other seeds. Additionally, in January, February and March, a larger variation can be seen compared to the rest of months (within the boxplots, the AUC values are more dispersed). This might be explained by the fact that during these months the data available was less, as winter surveillance is not usually carried out due to the low temperatures. As not much data is available, the predictions made on a small test set would be highly dependent on how the data is split.

Question 2: Which method has the best performance in terms of AUC score, considering also the possible variation in the predicted AUC for different data partitions and months?

The aim was to select a single balancing method that could be used for all months. I decided to choose a balancing method based not only on the highest AUC but also based on the variation in AUC found for the ten different partitions, as we also aimed for a methodology that could provide robust results, and not being affected by the random individual data partitions. Therefore, I looked for a metric that simultaneously considered a mean AUC and a low variation of the different AUCs.

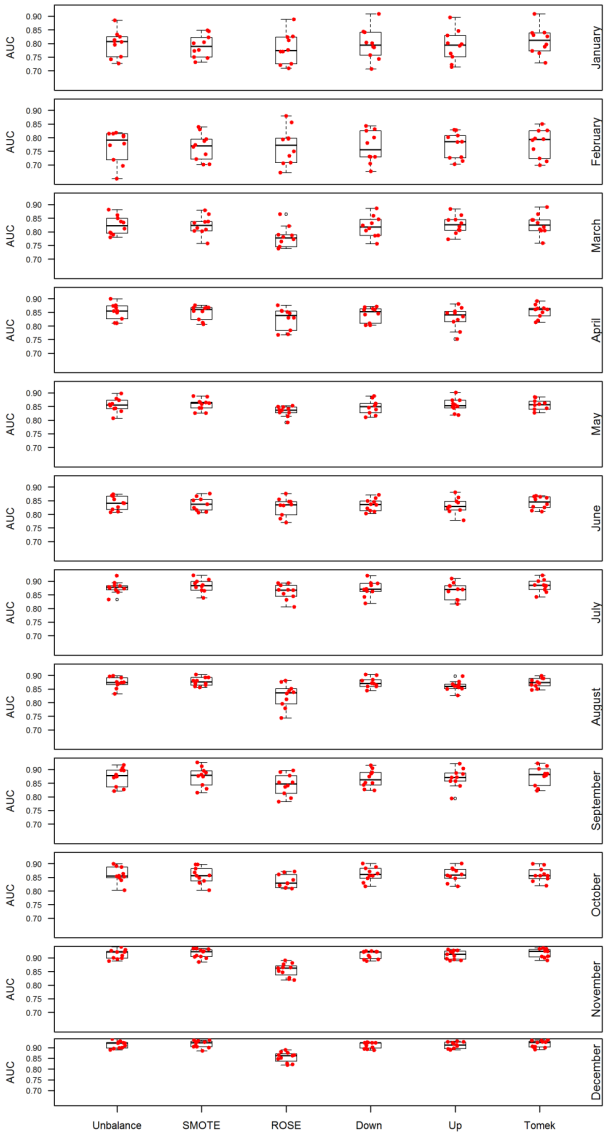


Figure 3.13: AUC scores calculated for each of the ten data partitions (red dots), plotted by month and for each balancing method. AUC values for each data partition are displayed on the Y-axis. For each method within each month, boxplots show the distribution of the AUC.

Method: I calculated the signal-to-noise ratio, which is the reciprocal to the coefficient of variation. The signal-to-noise ratio is defined as ratio between the mean to the standard deviation. Using the signal to noise ratio the 10 AUC values from each boxplot was used to calculate the mean

AUC and the standard deviation. This step resulted in a single value for each method: the higher the signal-to-noise ratio, the better the method performed among the months, as it shows a higher mean AUC value and a lower variation within a month and between them (mean/ sqrt(var)).

Results: The signal-to-noise ratio for each method was (Table 3.1):

Method:	**Unbalanced**	**SMOTE**	**ROSE**	**Down**	**Up**	**Tomek**
coefficient	15.55	15.08	13.60	15.53	17.29	15.75

Table 3.1:

The Up scaling method had the highest signal-to-noise ratio, meaning that it had the best combination of a high AUC and a low variance. In decreasing order, it was followed by Tomek, Unbalanced, Down scaling, SMOTE and lastly ROSE.

Conclusion: Up-scaling had the highest signal-to-noise ratio among the different balancing methods and unbalanced data.

Question 3: Question 3: Is the model using best balancing method (Up scaling) statistically different from a model using an imbalanced training dataset?

Method: To calculate confidence intervals and conduct inference, we simulated 10.000 values of the signal-to-noise ratio, by simulating normally distributed values within each month and for each method, with mean and variance as estimated from the 10 data splits. From these simulations, I extracted 10.000 signal-to-noise ratios for each method. Confidence intervals were then estimated as empirical intervals and p-values were estimated from the 10.000 simulations.

Results: The resulting p-value was p=0.36, indicating that there was not a significant difference between the apparently optimal method, “Up scaling” and the method “unbalanced” based on our data. However, because of the limited available data (only 10 different seeds tested), I assess the power of this comparison as low.

Conclusion: There was no significant difference between using the

Up scaling method or using none (just the imbalanced training set).

General conclusions

Considering the visual analysis of the density plots, which compares the different balancing methods using 6 different data partitions shown previously, I conclude that the distribution of the two classes, according to the probability of presence predicted by the models, did not result in a large variation among the different balancing methods.

Nevertheless, there was variation in the AUC calculated for the different methods across the different months. For January, February and March, no matter the method used, there was high variation among the seeds. A possible explanation for this behaviour is that the datasets was highly imbalanced, leading to the creation of a test set containing only few farms with the minority class. Thus, model performance would depend on which observations constitute the test set. If class imbalance is strong, with a small proportion of the minority class, then the test set would hold a smaller number of the minority class. This could make the results highly unstable i.e. different results might be obtained depending on which samples are found within the test set. Thus, an external validation may not be appropriate as the number of samples should be larger, to allow the test set to have enough samples of the minority class. Cross validation could have been used here but it was decided to test model performance on an independent test set, as model performance using cross validation was too optimistic compared to external validation (data not shown).

Even through there was no significant difference between RF model performances using the Up scaling method and the original imbalanced data set, I still chose to balance our training data set. I based this decision on the literature available stating that is recommendable to perform class balancing [Chawla et al., 2002, Japkowicz and Stephen, 2002, Kuhn and Johnson, 2013, Prati et al., 2004, Weiss and Provost, 2001]. On the other hand, the high p-value obtained here, might be the result of a low number of seeds evaluated and this result could be improved by increasing the test to a higher number of seeds. I did not try this due to the very large computational time required.

Final remarks

In this section:

1. Model performance comparison was made between the different methods: I used the AUC to compare the performance of different balancing methods by predicting on the same test set. For a single month, for example January, and for a given seed (for instance seed 1), the AUC values are comparable because they were calculated using the same dataset (the set for January).

2. The aim was not to compare the AUC from different months but to compare the results from different balancing methods, and the effect of the different data partitions per each individual month. However, AUCs for different months were used in the calculation of the single signal to noise ratio for each method. This was made only for the practical reason of using only one balancing method to run models using data from any given month.

AUC value can be used not only to compare the performance of different models, but also to assess the model performance of a single model on a given dataset. AUC is a metric that can be used to determine the predictive power of a classifier only by looking at its value: for example, a model with a AUC = 0.95 is consider to have an excellent performance, while a model with a AUC = 0.6 is consider to be a poor classifier. This is because the AUC score is the probability that "a randomly chosen true positive sample will have a higher rank than a randomly chosen true negative sample" [Fawcett, 2006]. In other words, the AUC is the probability of correct ranking of a random "positive - negative" pair. In the second manuscript, the AUC calculated for each month is not used to compare the model performance between different months, but to expresses if the monthly models (in this case the map) can be trusted or not. It was not the objective to compare the model performance between different months.

The AUC value is used widely for estimating performance of different modelling techniques in ecological modelling [Elith and Leathwick, 2009, McPHERSON et al., 2004]. However, its use has been criticised [Lobo et al., 2008], especially when used as a single value that the ROC might provide rather than the e.g. the shape of the curve [Jiménez-Valverde et al., 2009]. Nevertheless, for the second manuscript, I decided to use AUC values mainly because it is a threshold independent

metric, contrary to sensitivity/specificity or accuracy. AUC is also insensitive to class distribution [Fawcett, 2006].

Evaluating the probability of absence in relation to the observed abundance

Species distribution models (SDM) are used to estimate the environmental suitability (probability of presence) from presence-absence data collected and the environmental features measured on those collection sites. The probability of presence indicates the likelihood of species occurrence based on environmental predictors and therefore, it represents how well the species ecological requirements are met [VanDerWal et al., 2009]. Thus it is expected that areas with high probability of presence will be areas of high abundance, as the local abundance should be driven by the same environmental factors that drive the suitability.

I examined the relation between the predicted probability of presence and the observed abundance for the Obsoletus and Pulicaris ensembles and for *C. imicola*. The hypothesis was that there is a positive relation relationship between probability of presence and abundance.

Methods

For each month, a map of the probability of presence was generated based on presence/absence data. I used the same monthly training and test sets used for producing the interpolation maps and the mean abundance maps shown in the additional file of Manuscript III. They were used because the training and test sets were already created and the mean abundance was already calculated.

Each of these monthly training and test set were created partitioning the data randomly, with the training set containing 70 % of the data and the test set containing remaining 30 %.

These training sets contained the mean abundance calculated per farm. To create probability of presence maps, it is necessary to have the data as presence/absence format and therefore, the monthly mean abundance was transformed into presence/absence data as follow: for Obsoletus and the Pulicaris ensembles, the farms were classified as "Presence" if the mean abundance was equal or higher than 5 and for *C. imicola*, the farms were classified as "Presence" if the monthly mean

abundance was higher than 0. These thresholds were based on the European Commission thresholds used in their definition of *Culicoides* Presence or Absence. The farms that did not meet these requirements were classified as "Absence".

For each month, the training set was balanced using the Up-scaling method, which balances the class distribution of the training set by sampling with replacement from the minority class. The balanced training set was used for fitting the RF model and the resulting model was used to map the probability of presence using the 85 raster predictors (the same set of predictors used in Manuscript III).

The test set, which contained the mean abundance calculated per farm, was used to extract the probability of presence from the map. The predicted probability of presence was plotted against the observed abundance in order to visualise a possible relation between them.

The probability of presence map for the mean abundance data was only calculated to analyse the relation between environmental suitability and vector abundance.

Results

For the Obsoletus ensemble, there was a positive trend between vector abundance and probability of presence. For January, February and December the predicted probability of presence remained low, regardless of the observed abundance. For the rest of the months, the relation between the probability of presence and the abundance showed triangular shaped scatter plots, with the lower values of abundance being predicted at any range of the probability of presence, but as the abundance increases the predicted probability of presence increases too. In other words, high probability of presence is predicted for any abundance value, but high abundance is only observed at high probabilities of presence (Figure 3.14).

For the Pulicaris ensemble, the scatter plots between the probability of presence and the abundance showed a similar pattern as did for the Obsoletus ensemble, except for August and September, where no explicit positive relation can be observed (Figure 3.15).

For *C. imicola*, the relation between the predicted probability of presence and the abundance is less clear than for the Obsoletus and

the Pulicaris ensembles. In general, abundances values higher than zero where predicted mostly as high probability of presence (Figure 3.16).

Discussion

Probability of presence maps indicates how suitable an area is for the establishment of a given species or the probability of finding the species there and therefore, is an estimation of the optimal environmental conditions. These conditions are also considered to the same drivers for local abundance. Because of this, it is expected that abundance must be related to the probability of presence. Our results suggested that in general, there is a positive relation between the predicted probability of presence and the vector abundance. This is in agreement with the results obtained by VanDerWal et al. [2009], who analysed the environmental suitability and the abundance of tropical vertebrates in Australia. The scatter plots presented in their study showed a "wedge-shape point distribution in which the "upper limit of abundance increases at higher environmental suitability" [VanDerWal et al., 2009].

Other researchers have investigated the relationship between the environmental suitability and the abundance. A review can be found in Estrada and Arroyo [2012]. Not all attempts to find the this relation succeeded, for instance Jiménez-Valverde et al. [2009] failed to find a relationship between abundance and presence/absence models. This was the case for *C. imicola* and, in a lesser extent for the Pulicaris ensemble.

Conclusion

The results showed that probability of presence cannot provide a direct value of mean abundance, but instead can be used to estimate the upper limit of the abundance. In other words, in areas with a high probability of presence, any abundance value can be expected (from null to the maximum abundance) but in areas with low probability of presence, low values of abundance are expected.

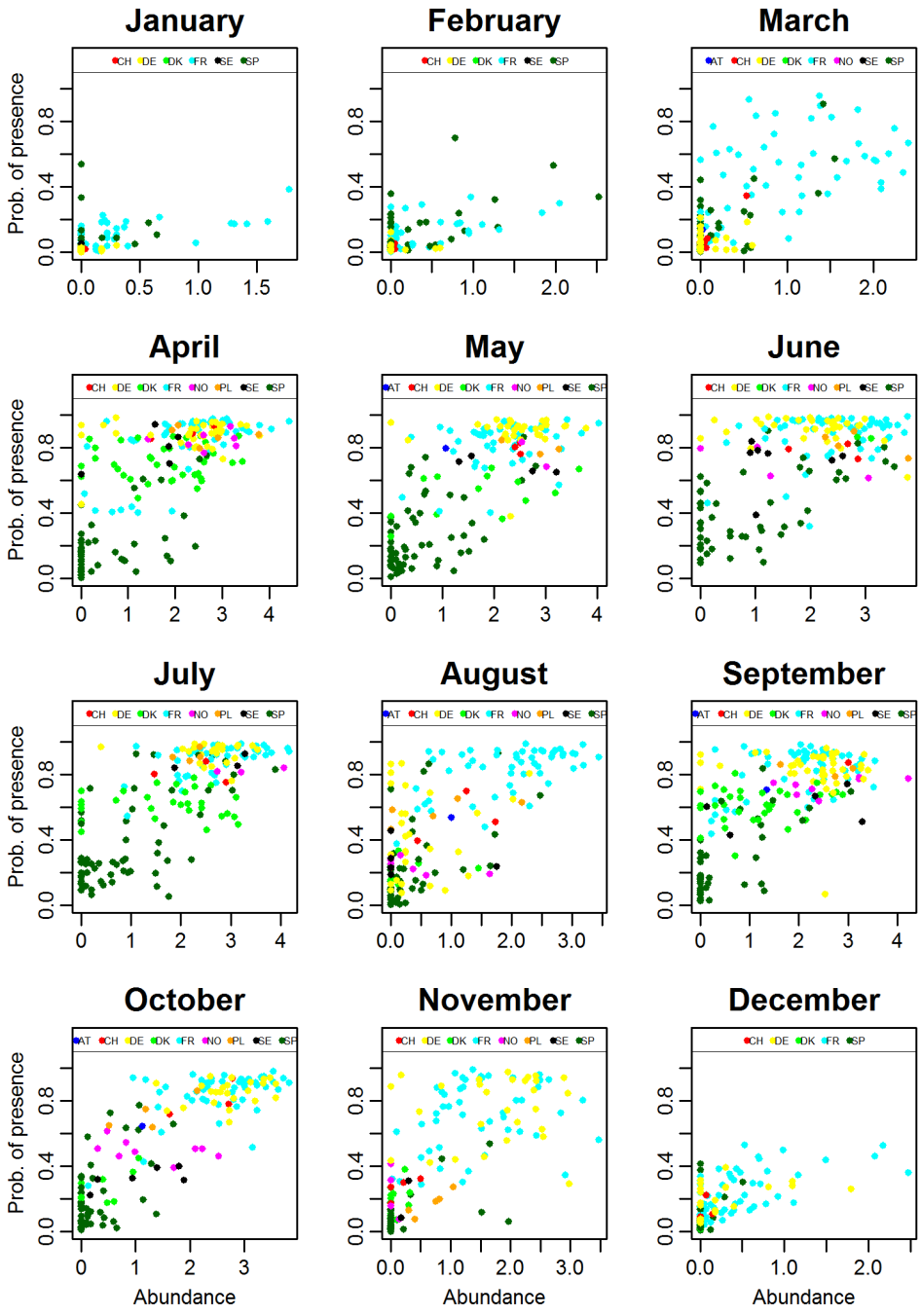


Figure 3.14: Monthly scatter plots of the predicted probability of presence (Y- axis) and the observed Obsoletus ensemble abundance ($\log_{10}$ scaled) (X-axis).

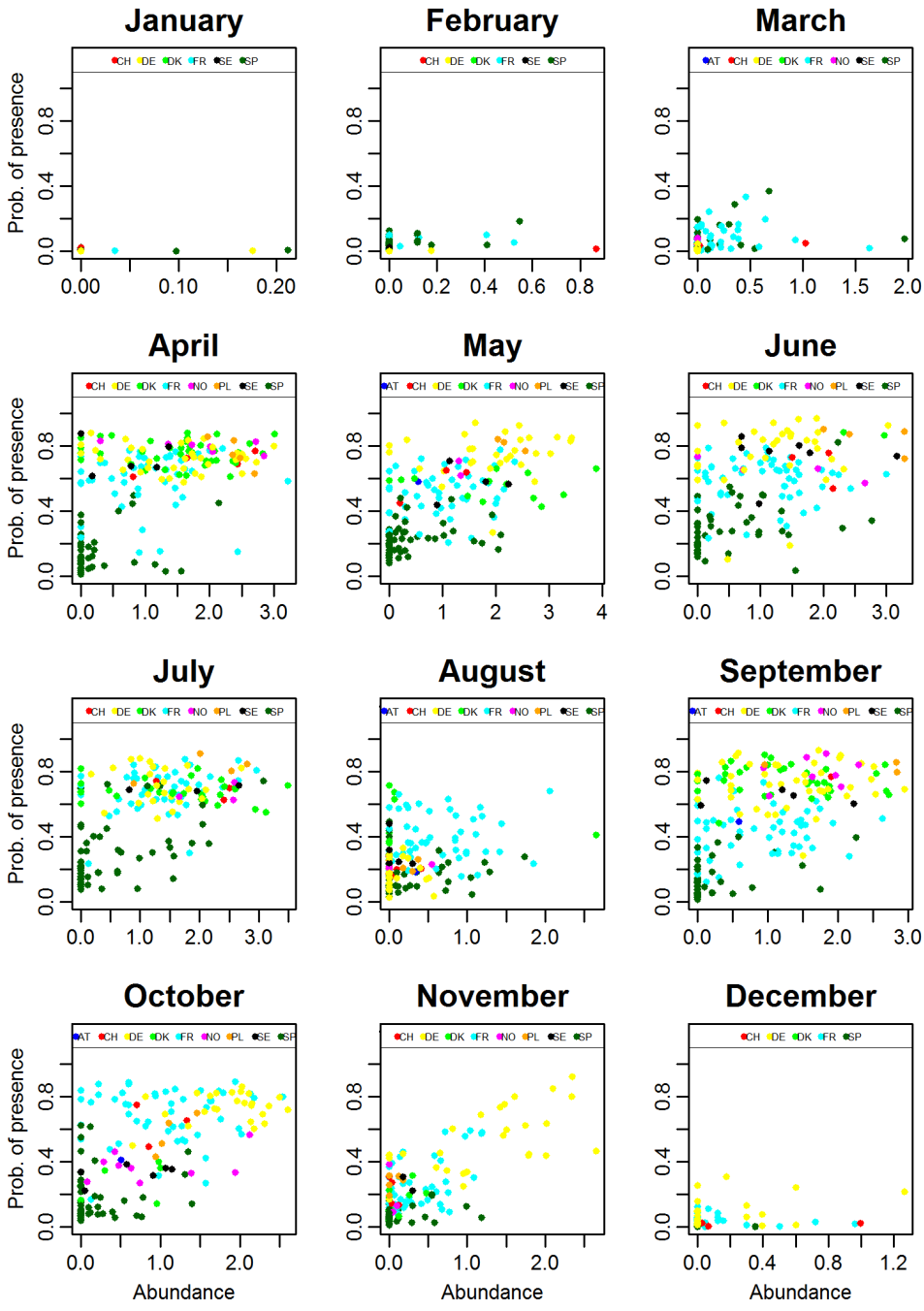


Figure 3.15: Monthly scatter plots of the predicted probability of presence and the observed Pulicaris ensemble abundance ($\log_{10}$ scaled) (X-axis).

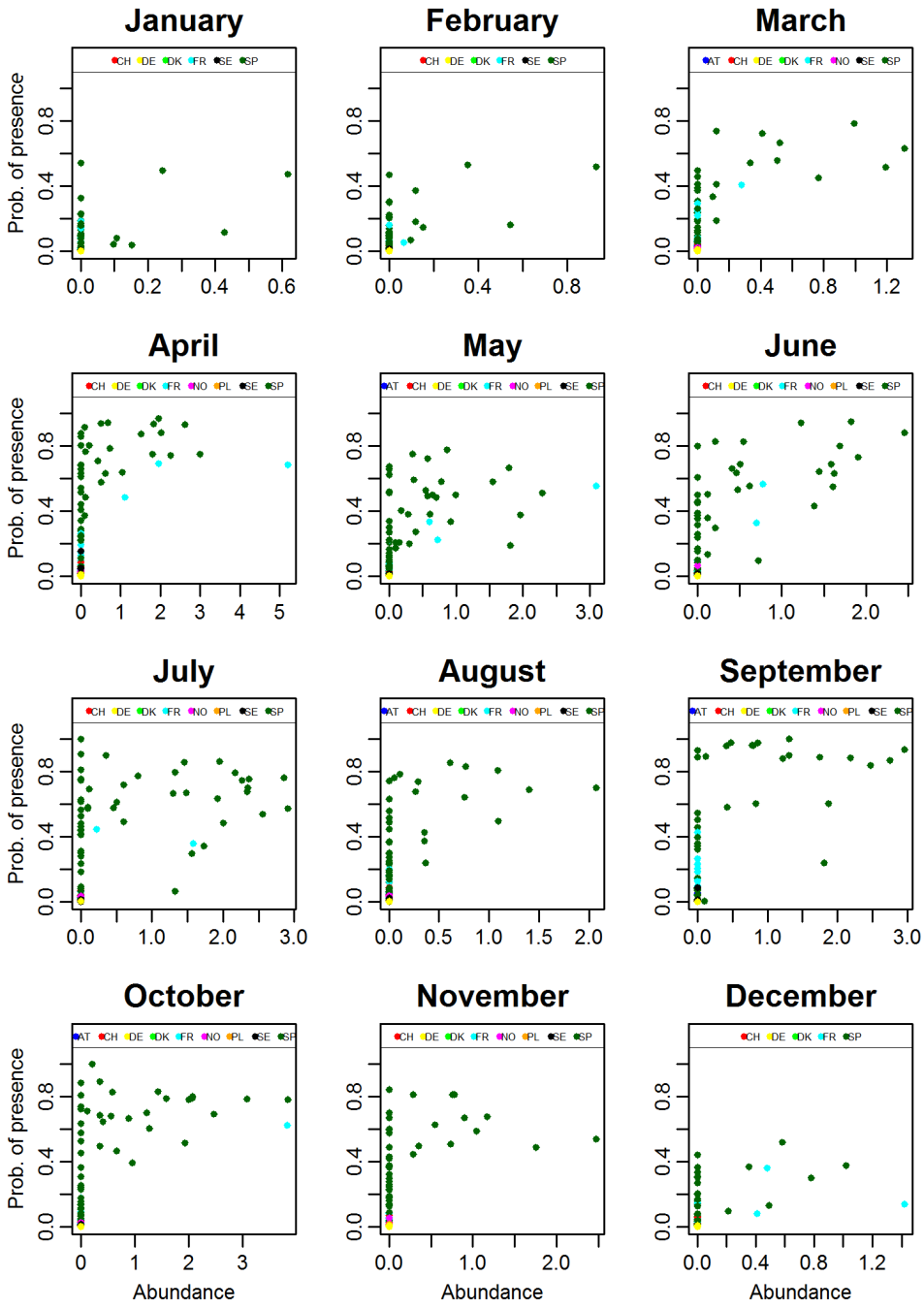


Figure 3.16: Monthly scatter plots of the predicted probability of presence and the observed *C. imicola* abundance ($\log_{10}$ scaled) (X-axis).

3.3 Manuscript III

Modelling the monthly abundance of *Culicoides* biting midges in nine European countries using Random Forest machine learning technique

Ana Carolina Cuéllar, Lene Jung Kjær, Andreas Baum, Anders Stockmarr, Henrik Skovgard, Søren Achim Nielsen, Mats Gunnar Andersson, Anders Lindström, Jan Chirico, Renke Lühken, Sonja Steinke, Ellen Kiel, Jörn Gethmann, Franz J. Conraths, Magdalena Larska, Marcin Smreczak, Anna Orłowska, Inger Hamnes, Ståle Sviland, Petter Hopp, Katharina Brugger, Franz Rubel, Thomas Balenghien, Claire Garros, Ignace Rakotoarivony, Xavier Allène, Jonathan Lhoir, David Chavernac, Jean-Claude Delécolle, Bruno Mathieu, Delphine Delécolle, Marie-Laure Setier-Rio, Roger Venail, Bethsabée Scheid, Miguel Ángel Miranda Chueca, Carlos Barceló, Javier Lucientes, Rosa Estrada, Alexander Mathis, Wesley Tack and Rene Bødker.

Manuscript in preparation.

Modelling the monthly abundance of *Culicoides* biting midges in nine European countries using Random Forest machine learning technique

Ana Carolina Cuéllar[1], Lene Jung Kjær[1], Andreas Baum[2], Anders Stockmarr[2], Henrik Skovgard[3], Søren Achim Nielsen[4], Mats Gunnar Andersson[5], Anders Lindström[5], Jan Chirico[5], Renke Lühken[6], Sonja Steinke[7], Ellen Kiel[7], Jörn Gethmann[8], Franz J. Conraths[8], Magdalena Larska[9], Marcin Smreczak[9],Anna Orłowska[9], Inger Hamnes[10], Ståle Sviland[10], Petter Hopp[10], Katharina Brugger[11], Franz Rubel[11], Thomas Balenghien[12], Claire Garros[12], Ignace Rakotoarivony[12], Xavier Allène[12], Jonathan Lhoir[12], David Chavernac[12], Jean-Claude Delécolle[13], Bruno Mathieu[13], Delphine Delécolle[13], Marie-Laure Setier-Rio[14], Roger Venail[14,18], Bethsabée Scheid[14], Miguel Ángel Miranda Chueca[15], Carlos Barceló[15], Javier Lucientes[16], Rosa Estrada[16], Alexander Mathis[17], Wesley Tack[18] and Rene Bødker[1]

[1]Division for Diagnostics and Scientific Advice, National Veterinary Institute, Technical University of Denmark (DTU), Lyngby, Denmark [2]Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Lyngby, Denmark [3]Department of Agroecology - Entomology and Plant Pathology, Aarhus University, Aarhus, Denmark [4]Department of Science and Environment, Roskilde University, Roskilde, Denmark [5]National Veterinary Institute (SVA), Uppsala, Sweden [6]Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Reference and Research National Reference Centre for Tropical Infectious Diseases, Hamburg, Germany [7]Department of Biology and Environmental Sciences. Carl von Ossietzky University, Oldenburg, Germany [8]Institute of Epidemiology, Friedrich Loeffler Institute, Greifswald, Germany [9]Department of Virology, National Veterinary Research Institute, Pulawy, Poland [10]Norwegian Veterinary Institute, Oslo, Norway 11Institute for Veterinary Public Health, Vetmeduni, Vienna, Austria [12]CIRAD, UMR ASTRE, F-34398 Montpellier, France [13]Institute of parasitology and tropical pathology of Strasbourg, EA7292, Université de Strasbourg, Strasbourg, France [14]EID Méditerranée, Montpellier, France [15]Laboratory of Zoology, University of the Balearic Islands, Palma, Spain [16]Department of Animal Pathology, University of Zaragoza, Zaragoza, Spain [17]Institute of Parasitology, University of Zürich, Zürich, Switzerland, [18]Avia-GIS NV, Zoersel, Belgium.

Corresponding author, Ana Carolina Cuéllar: anacu@vet.dtu.dk

Abstract

Background: *Culicoides* biting midges may transmit virus to ruminant livestock causing diseases, such as bluetongue, Schmallenberg and African horse sickness. In the past decades, these diseases lead to important economic loses for farmers in Europe. As for any vector borne disease, vector abundance is a key factor for determining the risk of disease spread, and thus there is a need to predict the abundance of *Culicoides* species implied in the transmission of bluetongue and Schmallenberg. The objective of this work was to model and map the monthly abundance of *Culicoides* in Europe.

Methods: entomological data was obtained from 904 farms in nine European countries (Spain, France, Germany, Switzerland, Austria, Poland, Denmark, Sweden and Norway) from 2007 to 2013. Using environmental and climatic predictors from satellite imagery and the machine learning technique Random Forest, we predicted the average monthly abundance at a 1 km^2 resolution. We used external validation to assess model performance.

Results: The predictive power of the resulting models varied according to month and the *Culicoides* species ensembles predicted. Model performance was lower for winter months. Performance was higher for the Obsoletus ensemble, followed by the Pulicaris ensemble. *Culicoides imicola* had poor model performance. Distribution and abundance patterns corresponded well to the known distribution in Europe. The Random Forest model was able to distinguish differences in abundance between countries but was not able to predict vector abundance at individual farm scale.

Conclusion: The maps presented here constitute the first of their kind for Europe and can be used as essential inputs for R_0 modelling of present and future *Culicoides* borne infections.

Background

Biting midges of the genus *Culicoides* (Diptera: Ceratopogonidae) are small blood sucking insects responsible for the transmission of bluetongue, Schmallenberg and African horse sickness viruses to wild and domestic ruminants such as cattle, sheep and goats and to equids [1].

Within Europe, outbreaks of bluetongue (BT) and Schmallenberg diseases have caused large economic losses to farmers and to national and European veterinary authorities during the last decades [2,3]. In Europe, the occurrence of BT was previously restricted to countries of the Mediterranean Basin where *Culicoides imicola* Kieffer was implicated in the transmission of the bluetongue virus (BTV) [4]. BT was never reported in northern Europe until August 2006, when an unexpected BTV-8 outbreak started at the border of Belgium, Netherlands and Germany [5]. During the following years, BTV spread to north-western and central Europe [6–8]. To control the spread of bluetongue, the European Union imposed restrictions to animal movements in the affected countries, followed later by vaccination campaigns for ruminant livestock. In addition, extensive entomological surveillance programs were stablished within the member states in order to determine the *Culicoides* species composition and monitor vector seasonal dynamics [9]. Several entomological studies conducted in northern Europe at the time of the epidemic confirmed the absence of the Afro-Asian vector *C. imicola* and it became evident that the BT virus was transmitted by Palearctic and autochthonous species of *Culicoides* [10]. This was confirmed with the isolation of the BTV serotype 8 from wild specimens of *Culicoides obsoletus* (Meigen)/ *Culicoides scoticus* Downes & Kettle [11,12] , *Culicoides dewulfi* Goetghebuer [13] and *Culicoides chiopterus* (Meigen) [14,15].

The basic reproduction number R_0 represents the progression of a disease over time (i.e. whether transmission will fade out or grow epidemically). This value expresses the number of new cases generated from a single case, given the introduction of a pathogen into a naïve population [16,17]. The predicted basic reproduction number R_0 can be used to identify areas at high risk of disease outbreaks and therefore for the allocation of financial

resources and the establishment of control measurements by veterinarian authorities. For vector-borne diseases, different parameters are used to calculate R_0 such as vector density, which increases the number of bites a host would receive [16]. Thus, estimating the vector abundance spatially constitute a fundamental step to map the potential risk of disease (i.e. R_0 values) [17]. Vector abundance data collected on farms may be predicted to un-sampled areas in order to generate a continuous abundance surface [18] using statistical models and environmental and climatic features as predictors.

The geographical distribution of insect vectors is mainly driven by environmental and climatic factors that influence their ecology [19]. Lower temperatures delimitate the geographical range of most insect species, and within the optimal temperature range, warmer temperatures decrease the length of their life cycle stages, thus increasing the number of generations produced per season [20–22]. In addition, precipitation influences the abundance of adult insect populations as it favours the persistence of optimal breeding sites for some species of biting midges [23], such as *C. imicola* that breeds in wet and organic enriched soils and mud [24]. Precipitation is also related to humidity, a variable that affects the level of activity and survival of adult *Culicoides* specimens preventing them from desiccating [20,22,25]. Because of this strong association between environmental factors and insect vector abundance, satellite-derived environmental variables, used as surrogates of ground measurements (for instance from weather stations), can be used as spatial predictors for vector abundance. Remote sensed imagery constitutes an important source of environmental information and has been used worldwide and repeatedly in the last decades for mapping vector distribution and disease risk [26].

Many of the available distribution maps for *C. imicola* and Obsoletus in Europe show the distribution as the probability of presence, as a result of modelling Presence versus Absence data [27–34]. However, *Culicoides* abundance models can also be found in the literature, either at a national scale [34–36] or at a continental scale for *C. imicola* [32,37] and for the Obsoletus group [38,39]. The *Culicoides* maps available for Europe, at a continental scale, are usually made with abundance data collected within a limited area of the entire region and, after modelling, the response is extrapolated to the rest of the

continent, beyond the domain of the sampled farms. Tatem [32] used *C. imicola* abundance collected in 87 sites across Portugal and extrapolated the abundance to the rest of Europe. Similarly, Baylis [37] used data from 49 sites from Portugal, Spain and Morocco and predicted the abundance in the Mediterranean basin. A major limitation of extrapolation is that predictions are made beyond the environmental range in which data collections were taken and therefore, based on the assumption that the environmental relationship with abundance follows a simple function [40]. In addition, validating the model predictions of these areas is only possible if there are external vector data. As an example, Pili et al [41] showed that a previous *C. imicola* abundance map made for the Mediterranean basin, using sampled farms from Sicily [29], seemed to be inaccurate at predicting vector abundance in Sardinia. Therefore the existing distribution and abundance maps of *Culicoides* for Europe based on extrapolation provide basic knowledge of the geographical distribution and a general overview about the vector abundance across Europe, but they should be interpreted with care by disease control decision makers. Machine learning techniques are algorithms that, like the classical statistical models, can be used to predict an outcome using predictor variables. The machine learning technique Random Forest (RF) has proven to outperform classical approaches for species distribution modelling [30,42,43]. We here hypothesised that *Culicoides* abundance may be predicted for a large area of Europe, using a RF approach on entomological data collected in farms across nine European countries and Fourier transformed satellite imagery of 1km^2 resolution containing environmental, climatic and land cover predictors, which have shown to have an effect in previous *Culicoides* studies [34,43–45]. The data set used is the largest entomological data set aggregated to date by collaboration of nine European countries. We furthermore hypothesised that this data set and predictors may be used for predicting *Culicoides* abundance at farm scale, as we consider a resolution of 1km^2 to represent the interaction between environment and abundance. We predicted the average monthly abundance for each year sampled, within an area covering transects from southern Spain to northern Scandinavia. We generated an average abundance map per month for Obsoletus and Pulicaris ensembles and for *C. imicola*. These resulting maps may be used as inputs for future R_0 models of *Culicoides*-borne diseases and other risk assessment modelling.

Methods

Culicoides dataset

Culicoides data was collected from cattle, sheep and horse farms in Spain, France, Germany, Austria, Switzerland, Denmark, Norway, Sweden and Poland from 2007 to 2013 [46,47] (figure 1).

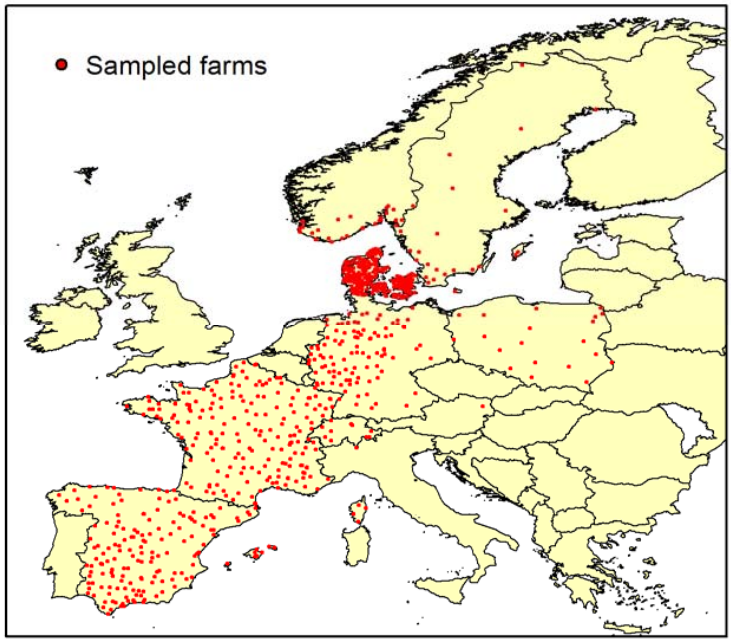


Figure 1. Available data from sampled farms in Europe during entomological surveys from 2007 to 2013. (taken and modified from Cuéllar et al. 2018 [46])

Black light suction traps were placed outside each farm and they were usually operational once a week during the sampling period. Specimens were identified to species level for *C. imicola* and aggregated when they belonged to the Obsoletus and Pulicaris ensemble. We use the term "ensemble" to refer to a group of sympatric species for which morphological identification is sometimes not possible or difficult, and without phylogenetic meaning. This term has been used in two previous analyses using this same dataset [46,47]. The data

was originally collected as part of national surveillance programs in each country and therefore sampling protocols differed in sampling period and number of nights sampled and the trap type used. Catch data included farm ID, latitude, longitude, date of trapping, number of sampling nights, number of *Culicoides* caught per observation (differentiated into *C. imicola*, Obsoletus ensemble and Pulicaris ensemble) and type of trap used. The segregation of females by their parity stage was not made by all countries and therefore, we did not differentiate regarding their gonotrophic stage. All countries set only one trap per farm with the exception of Germany that placed three traps per farm. In order to retain only one observation per farm, we calculated the median abundance value from the German farms. We calculated conversion factors for the BG sentinel and mini CDC traps in order to make catch data comparable to the data obtained from the Ondersterpoort traps. Details on the sampling protocol and conversion factors can be found in Cuéllar et al, 2018 [46].

Predictor variables

We used environmental and climatic data together with estimates of production animal density and land cover features as predictor variables of the biting midge abundance. All the predictors were in a 1x1 km raster format.

Environmental predictors were derived from a MODIS temporal series from 2001 to 2012. We considered Mid-infrared (MIR), daytime Land Surface Temperature (dLST), night-time Land Surface Temperature (nLST), Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) and each variable had been processed applying a Temporal Fourier Analysis (TFA) [48]. This technique fits a sine function to the observations taken at each time interval of the temporal series and decomposes it into its harmonic components and other descriptors of a temporal series such as maximum, minimum and mean. This process results in 14 new raster images for every environmental variable (Table 1). This dataset was created and processed by the TALA research group of Oxford University and obtained through the EDENext project [49].

Fourier Component	Description
A0	Fourier mean for the entire time series
A1	Amplitude of annual cycle
A2	Amplitude of bi-annual cycle
A3	Amplitude of tri-annual cycle
P1	Phase of annual cycle
P2	Phase of bi-annual cycle
P3	Phase of tri-annual cycle
DA	Proportion of total variance due to all three cycles
D1	Proportion of total variance due to annual cycle
D2	Proportion of total variance due to bi-annual cycle
D3	Proportion of total variance due to tri-annual cycle
MN	Minimum value
MX	Maximum value
VR	Total variance

Table 1. Products of Temporal Fourier Analysis obtained from each of the five? remote sensing variables.

The Bioclim raster dataset was obtained from Worldclim online database [50] and density animal data for cattle, chicken, goats, small ruminants and sheep were obtained from FAO "GeoNetwork" [51] (Table2).

Source	Code	Description
Modis (Fourier transformed) 2001-2012	MIR	Mid-infrared
	dLST	Daytime land surface temperature
	nLST	Nighttime land surface temperature
	NDVI	Normalized difference vegetation index
	EVI	Enhanced vegetation index
Bioclim 1960-1990	BIO 1	Annual mean temperature
	BIO 2	Mean diurnal range: mean of monthly (max. temp - min. temp)
	BIO 3	Isothermality (BIO2/BIO7) (*100)
	BIO 4	Temperature seasonality (standard deviation *100)
	BIO 5	Max. temperature of warmest month
	BIO 6	Min. temperature of coldest month
	BIO 7	Temperature annual range (BIO5-BIO6)
	BIO 8	Mean temperature of wettest quarter
	BIO 9	Mean temperature of driest quarter
	BIO 10	Mean temperature of warmest quarter
	BIO 11	Mean temperature of coldest quarter
	BIO 12	Annual precipitation
	BIO 13	Precipitation of wettest month
	BIO 14	Precipitation of driest month
	BIO 15	Precipitation seasonality (coefficient of variation)
	BIO 16	Precipitation of wettest quarter
	BIO 17	Precipitation of driest quarter
	BIO 18	Precipitation of warmest quarter
	BIO 19	Precipitation of coldest quarter
	Altitude	Digital elevation model (DEM)

Corine Land Cover	CLC 12	Non-irrigated arable land
	CLC 13	Permanently irrigated land
	CLC 15-17	Vineyards, fruit trees and berry plantations, olive groves
	CLC 18	Pastures
	CLC 19	Annual crops associated with permanent crops
	CLC 20	Complex cultivation patterns
	CLC 21	Land principally occupied by agriculture with significant areas of natural vegetation
	CLC 22	Agro-forestry areas
	CLC 23	Broad-leaved forest
	CLC 24	Coniferous forest
	CLC 25	Mixed forest
	CLC 26	Natural grasslands
	CLC 29	Transitional woodland-shrub
	CLC 35	Inland marshes
	CLC 40	Water courses
	CLC 41	Water bodies

Table 2. Environmental and land cover predictors used to model the probability of *Culicoides* presence.

We used land cover in a resolution of 250 m from the CORINE Land Cover [52], from which we extracted 16 classes, we considered relevant for *Culicoides* occurrence. Each class was transformed into binary images according to the presence/absence of the class. From these binary images we calculated the number of pixels containing the class for every 1 km^2 pixel and created maps displaying the frequency of each class per pixel. Details on the processing method for CLC map are described in chapter 2 (Materials and Methods) of the thesis.

All the raster predictors need to cover the same geographic region (i.e, have the same extent) and contain the same number of pixels (rows and columns). Therefore, pre-processing of the raster files included re-projection to WGS 84 geographic system, cropping to fit the study area, resampling and extent matching. Pre-processing was done using the free software R 3.4.2 [53] (package raster [54]).

We visually analysed potential correlation among the predictors by plotting the variables against each other in pairwise combinations. We analysed the pairwise correlation within the Fourier transformed variables and within the Bioclim variables separately. We identified pairs of highly correlated variables and from each correlated pair we removed one of the variables from the analysis. In total we removed 25 predictors: BIO 4, BIO 5, BIO 6, BIO 10, BIO 11, BIO 12, BIO 16, BIO 17, BIO 9, Small Ruminants, MIR_MN, MIR_MX, dLST_MN, nLST_MX, nLST_MN, nLST_MX, NDVI_MN, NDVI_MX, EVI_MN, EVI_MX, MIR_VR, dLST_VR, dLST_D3, nLST_VR, nLST_D3,NDVI_VR and EVI_VR . It should be noticed that from the Fourier transformed predictors, all the "minimum" and "maximum" variables were highly correlated to the mean variables and were therefore removed. The "variance" predictor (VR) was very highly correlated to the annual amplitude "A1" and was also removed from the analysis.

As the monthly mean abundance of *Culicoides* showed large variation among the years (data not shown), we decided to include, in each monthly model, the year of sampling as a predictor variable. The variable "year" was added as a set of seven binary dummy variables (one for each year), each one containing values of 0 and 1, (1 being assigned to each dummy when the catches were made during that year). Including this variable as predictor allows the model to make predictions for each individual year separately. For each month, RF model was fitted and the resulting model was then used to a produce a prediction map for each individual year. To be able to generate a map, it is a requirement for the "raster" package in R that all the predictors used for training the model are indeed available in raster format. Thus, we created a dummy raster file for each of the seven dummy variables of "year" in a 1 km^2 resolution. The RF model was run seven times obtaining a map for each of the seven sampled years.

For each month, we used seven annual prediction maps to calculate: (i) the average predictions of the seven years and (i) the coefficient of variation as: Coef. Var = (standard deviation) / mean. These calculations were done per each pixel using the values corresponding to each year (N=7).

We considered this average map to be the best prediction of abundance in a future year and the coefficient of variation as measurement of the variation found within the seven year prediction. A standard deviation map has been created previously showing the variability for predictions made for *Culicoides impunctatus* in Scotland [45]. Instead of calculating the standard deviation, we here chose to calculate the coefficient of variation. The coefficient of variation is a way to calculate variation based on "mean units" and allows comparison of the variation of samples with different means. Because each pixel contained different mean values and, as the standard deviation is proportional to the mean value, pixels with a higher mean would exhibit higher standard deviation. Therefore, areas with higher abundance (higher mean) would risk leading to the wrong interpretation that the variation during the seven years were higher.

Modelling approach

We used the machine learning method "Random Forest" (RF) [55] to predict the abundance of biting midges to non-sampled areas. RF consists of an ensemble of decision trees (a forest) in which each tree contributes to a prediction of the new incoming observation. When the response variable is continuous, regression trees are created and the final prediction of a new sample is computed by averaging all the predictions from the different trees in the forest [56]. RF technique has previously been used to model the geographical distribution and/or abundance of vectors such as mosquitoes [43], biting midges [34,57] and non-arthropod vectors [58]. The advantages of using decision trees are that they can handle non-parametric data, they are robust to outliers and capable of identifying complex interactions between the response and the predictor variables. Additionally, RF ranks the most important predictors by calculating the improvement in the error when each variable is permuted [40,56].

We used the free software R 3.4.2 [53] (packages caret [59], randomForest [60] and raster [54]) to model and predict the abundance data using the above mentioned raster files as predictors. The caret package looks for the best number of candidate variables for splitting the data at each node (m_{try}) using a tuning grid. In this work, m_{try} has been set to 30 and the number of trees was set to 750 (ntree=750). We used a 5 fold cross validation for the tuning process.

Culicoides data management

The dataset was divided into 12 independent monthly subsets according to the month that each catch belonged to. For each monthly dataset, we first calculated the average abundance at each farm for each year sampled and log10 transformed it after adding 1.This resulted in 12 monthly data sets where each farm contained as many records as number of years sampled. These abundance estimates were considered as independent observations although they originated from the same farm.

Validation

We randomly divided each monthly dataset into a training and test set. The training set contained 70 % of the total farms sampled that month, and the test set contained the remaining 30 % of the farms. For each month, RF performance was analysed by predicting the test set observations (external validation) [45,59]. We plotted the predicted values as a function of the observed values of the test set observations. We used the Square Root Mean Error (RMSE) normalized with respect to the mean predicted value for each month (the normalised RMSE, i.e. nRMSE) in order compare the results from different months having different observed ranges. The lower the nRMSE, the better the model performance. Additionally, QQ plots were made to analyse if the residuals were normal distributed. A good model should generate normal distributed residuals.

Interpolation model

We wanted to determine if it would be feasible to use simple spatial interpolation to predict abundance at a large geographic scale and to compare it to RF modelling. Interpolation is a

simple method for predicting a response variable as it only depends on the spatial position and the value of the sampled locations and do not require additional predictor variables.

To carry out an interpolation, data points should contain only one record as interpolation fits a function between the data points. Therefore, we were not able to use the training datasets used previously for training the RF model, as the farms were containing averages from different years, and not a single record.

To compare interpolation and RF modelling, we decided to develop new RF models. We therefore calculated the monthly average per farm using the averages calculated previously per each year and used the overall mean abundance per each farm. This resulted in a dataset containing only one abundance estimate per farm. For each species we ran RF models on these 12 monthly datasets obtaining a single prediction map per month. For each species and month, a prediction map based on an interpolation approach was made using the same training dataset used for training the RF models retaining the test data set for evaluation. We used the Inverse Distance Weighted (IDW) algorithm, which makes predictions to un-sampled locations by calculating the average of the values measured at neighboring locations and gives higher weight to the points closer to the prediction point, with the weight decreasing as a function of the distance. IDW was used to predict the abundance of this same dataset in a previous analysis [46]. We used the IDW function (Geostatistical Analyst tool) in ArcMap 10.1 (ESRI, Redlands, CA, USA) using the following settings: power equal to 2, minimum neighbours equal to 10 and maximum neighbours equal to 15. The interpolation maps are the same as the ones produced previously by Cuéllar et al. [46].

To validate the interpolation maps based on the training datasets, we extracted the interpolated values for each farm of the test set and calculated the residuals (observed minus the predicted values in test dataset). We evaluated the interpolation performance by plotting the predicted values against the observed values and calculating the nRMSE.

Results

Model performance

The nRMSE for each month showed that, in general, RF performed well for the Obsoletus ensemble (nRMSE range= 0.45 - 2.56) and fairly well for the Pulicaris ensemble (nRMSE range= 0.71 – 3.07) but poorly for *C. imicola* (nRMSE range= 2.21 – 12.27). The performance of the RF models varied between the months with the nRMSE being higher than 1 during the colder months (table 3).

	Obsoletus ensemble	**Pulicaris ensemble**	***C. imicola***
Month	**nRMSE**	**nRMSE**	**nRMSE**
January	1.73	3.07	8.28
February	2.56	6.50	6.83
March	1.18	2.47	6.76
April	0.65	1.26	4.05
May	0.55	0.91	3.67
June	0.51	0.84	2.52
July	**0.45**	**0.71**	2.33
August	0.52	0.74	2.66
September	0.55	0.85	**2.21**
October	0.49	0.80	2.55
November	0.73	1.13	3.01
December	1.55	2.84	12.27

Table 3. nRMSE calculated for each month and each *Culicoides* ensemble/species. Bold letters show the lowest nRMSE.

In the general, there was a positive linear correlation between predicted and observed values for the Obsoletus ensemble. This trend was weakest in March, where a scatterplot of observed versus predicted values showed a cloud with a weak linear trend and a nRMSE of 1.18 (Figure 1a). The best model was for July with a nRMSE = 0.45 followed by October, with a nRMSE = 0.49. Analysing the predictions by country, the highest predictions were observed for Germany (May-November), followed by France while the lowest predictions were found for Spain. Considering each country separately, the monthly models were not able to predict the observed abundance very well. This can be observed from the fairly

horizontal patterns in the predictions for each country, despite a relatively larger variation in the observed abundance in each country (Figure 2a). For January, February, March, November and December (winter period), the QQ plots showed the residuals were not normal distributed (Figure 2b). In all the months, we found a high variation in the predictions from farms with null abundance; however this variation decreased as the observed abundance increased (figure 2a).

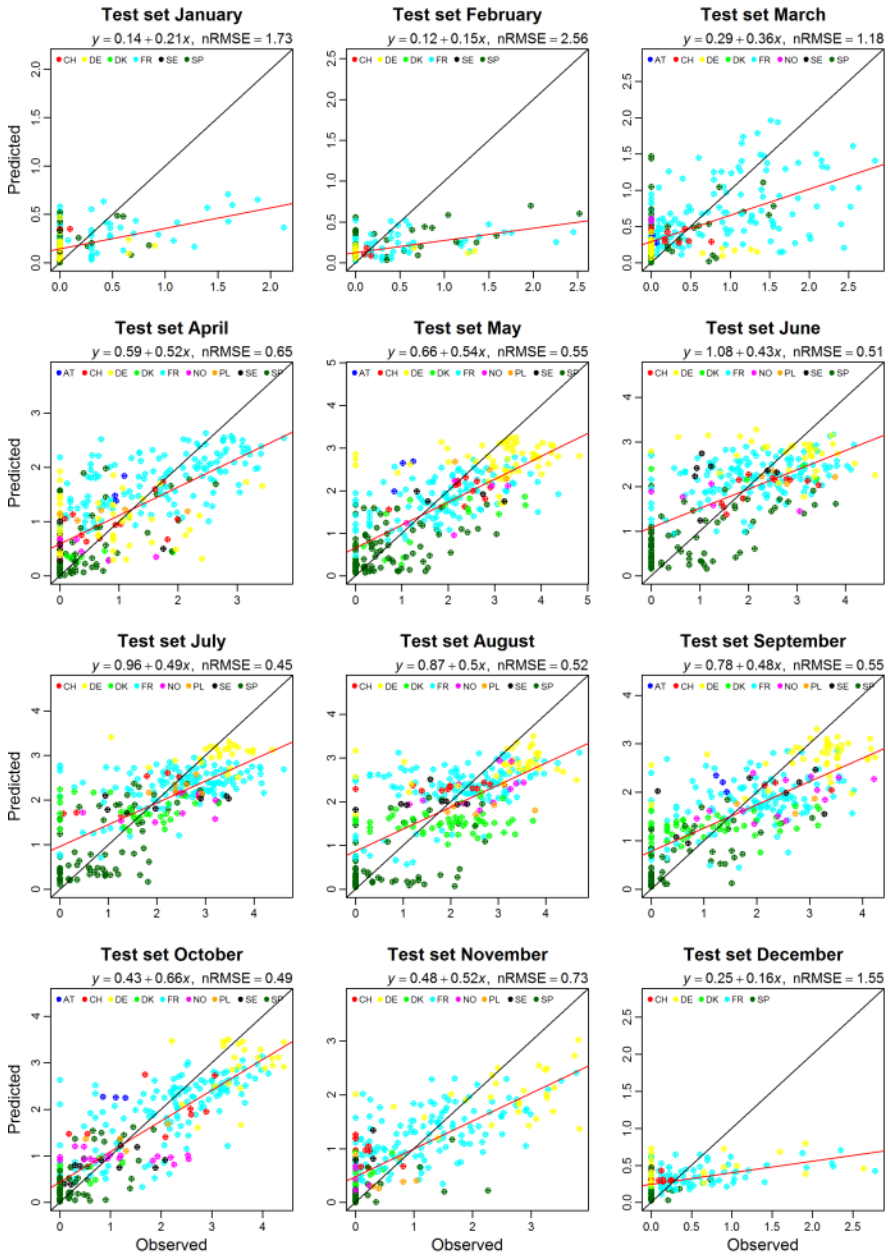


Figure 2 a. Scatter plot of the predicted and observed abundance of the Obsoletus ensemble. Red line: best linear model fit. Black line: perfect model fit. Note that scales depict log10 values and varies between different months.

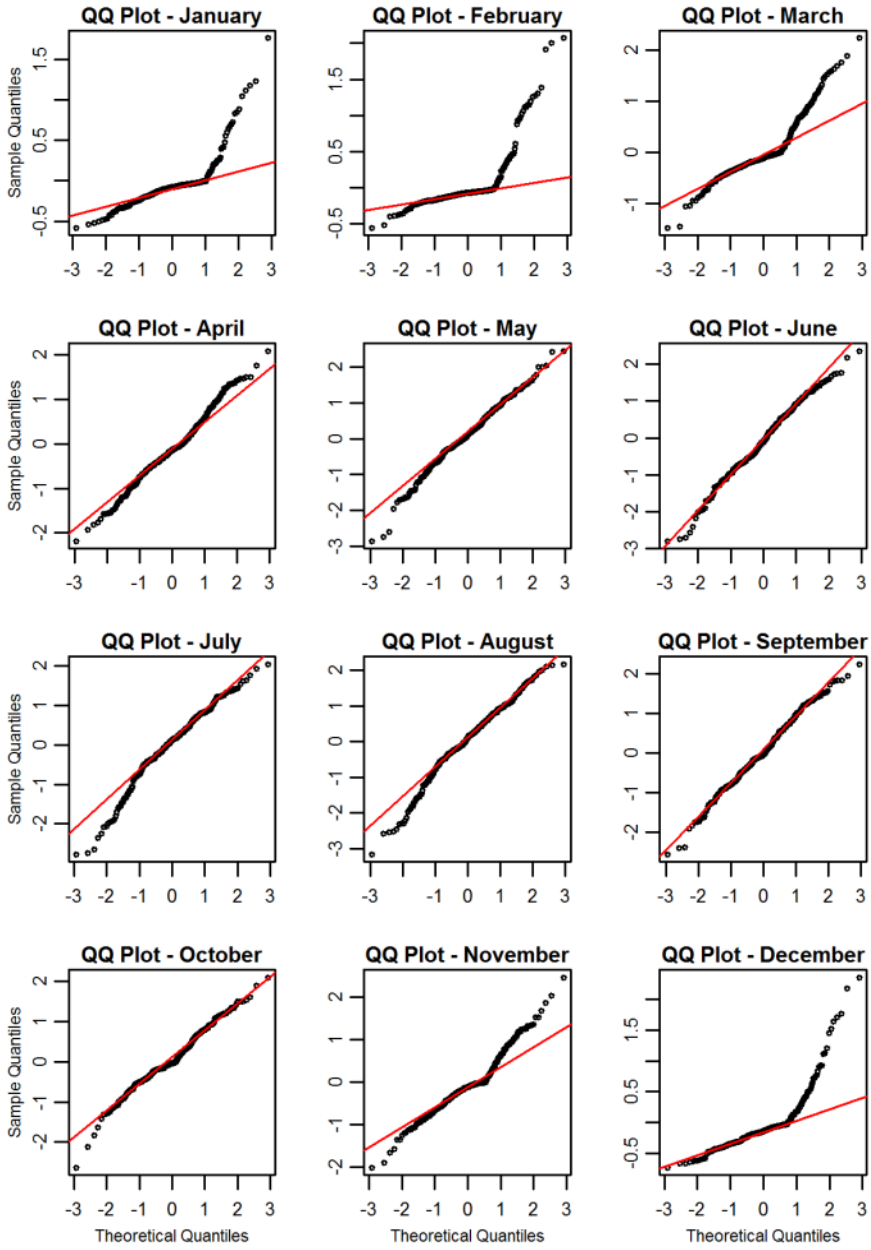


Figure 2 b. QQ plots of the residuals per month for the Obsoletus ensemble.

Performance of the Pulicaris ensemble model was poorer compared to the Obsoletus ensemble model, resulting in a minimum nRMSE of 0.71 in July (table 3). Again we found a poor ability of the model to predict the large observed variation between farms within each single country in the training dataset (Figure 3a). QQ plots of the residuals showed that the models from January, February, March, April, November and December were not normally distributed, indicating that the variance of the model predictions were not random but depended on the observed abundance (Figure 3b).

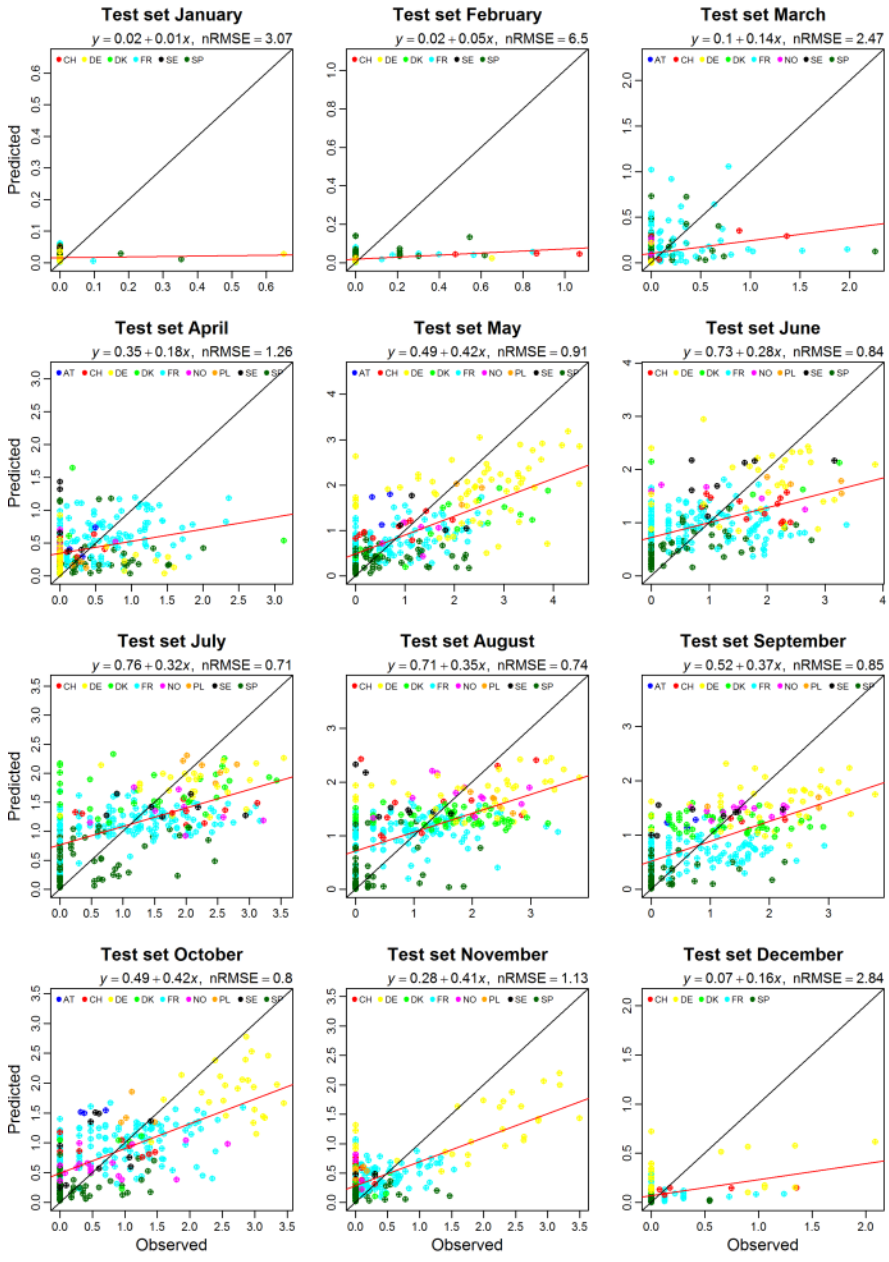


Figure 3 a. Scatter plot of the predicted and observed abundance of the Pulicaris ensemble. Red line: best linear model fit. Black line: perfect model fit.

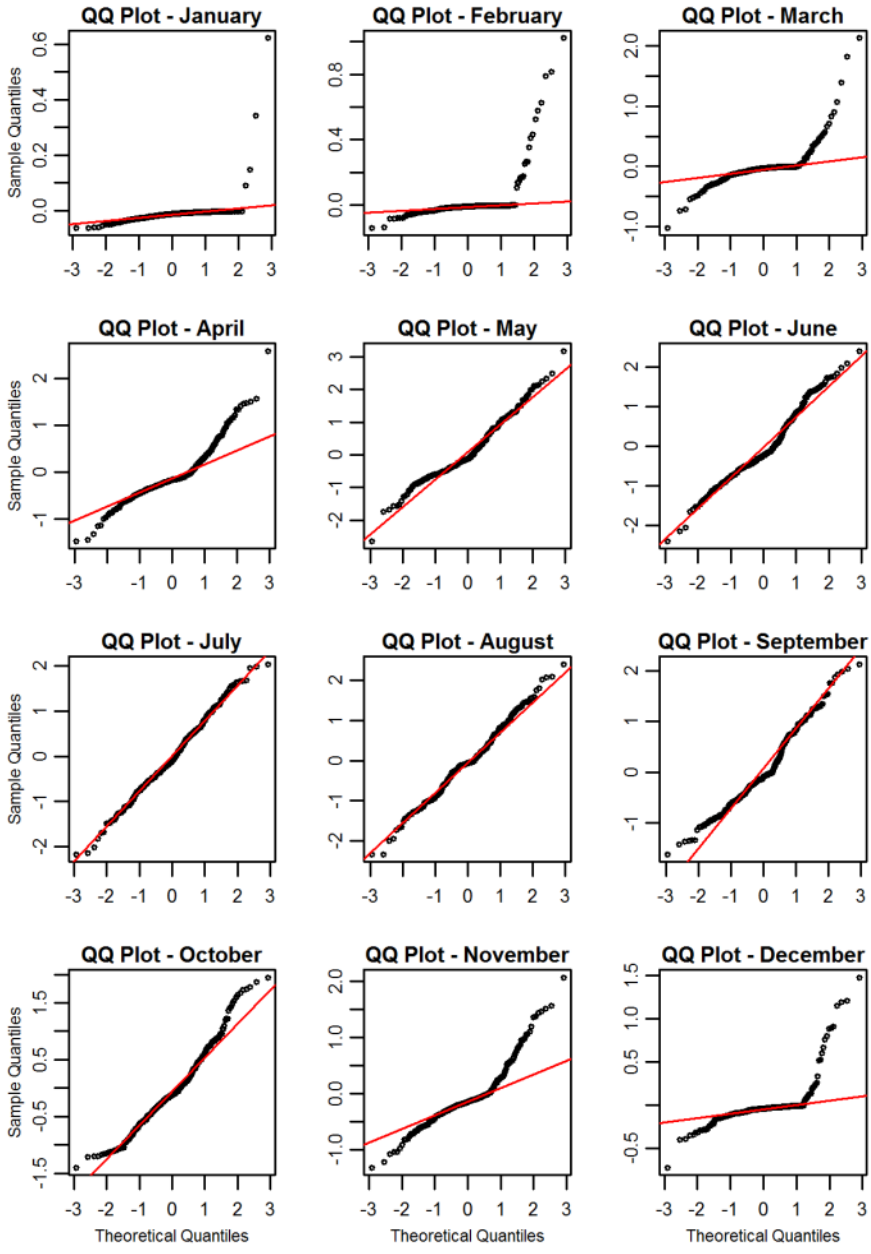


Figure 3 b. QQ plots of the residuals per month for the Pulicaris ensemble.

Performance of the *C. imicola* model was poor, as shown by the high nRMSE obtained for all the months (table 3). The best model obtained was for September, with the minimum nRMSE= 2.21 (figure 4a). The monthly models were incapable of predicting the highest observed abundances for *C. imicola*, resulting in similar predictions to those obtained for farms with observed null or low abundance (Figure 3a). The residuals were not normal distributed in any month (figure 4b).

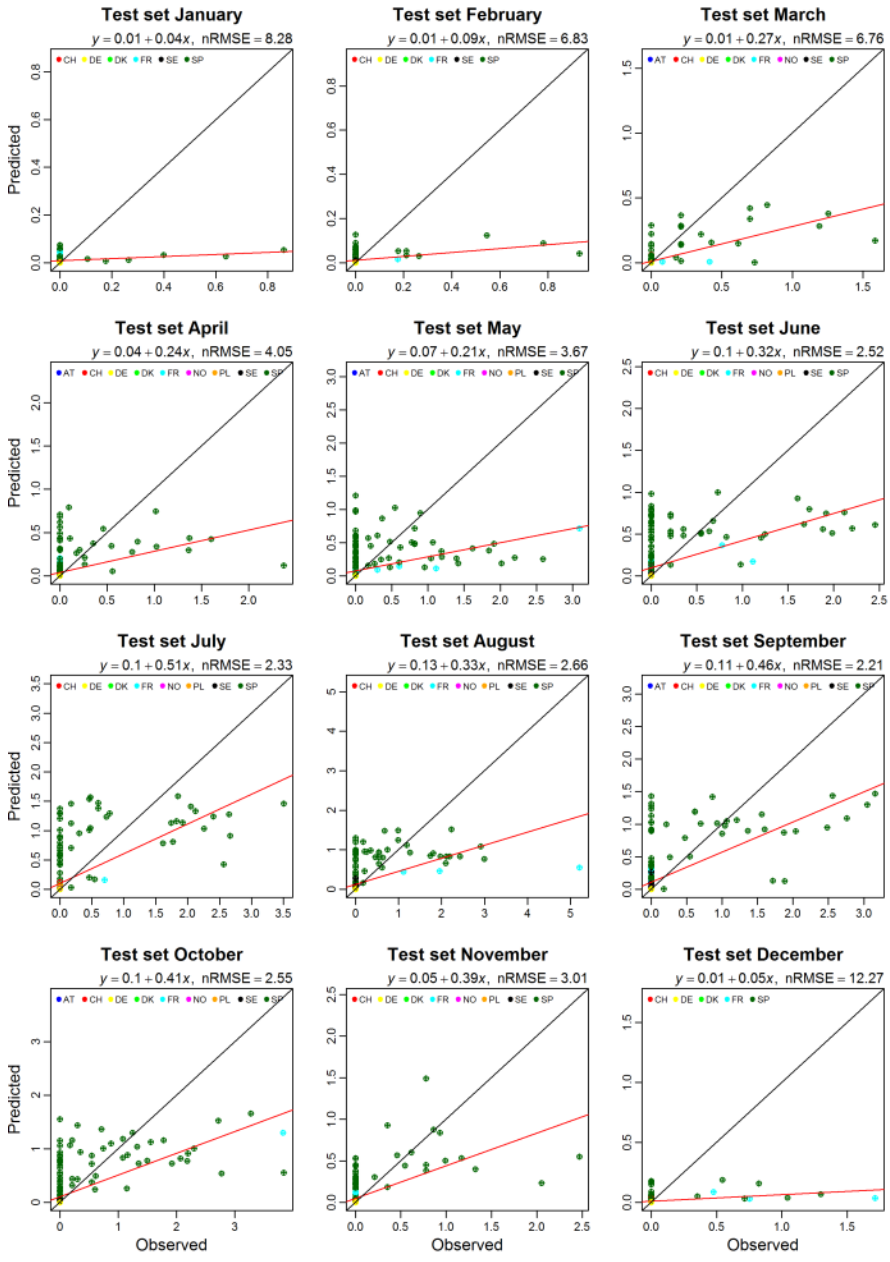


Figure 4 a. Scatter plot of the predicted and observed abundance of *C. imicola*. Red line: best linear model fit. Black line: perfect model fit.

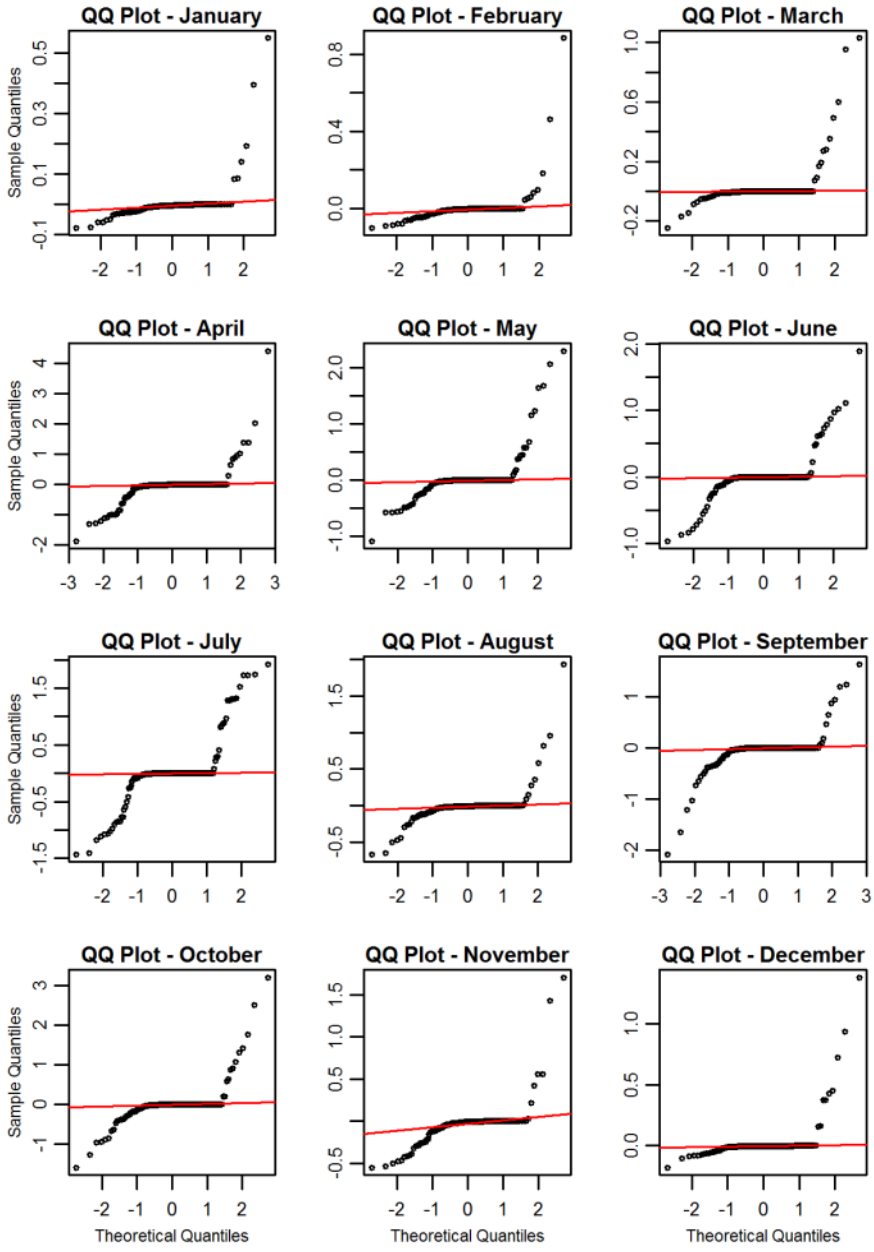


Figure 4 b. QQ plots of the residuals per month for *C. imicola*.

For the two ensembles and for *C. imicola*, the models consistently over-estimated observed low abundances and under-estimated observed high abundance as can be seen in figures 2a, 3a and 4a.

Average abundance maps from annual maps

The predicted abundance for the Obsoletus ensemble showed a seasonal pattern with very low abundance during January (<10 individuals) and February. During March, the predicted abundance of the Obsoletus ensemble started to increase in western France and along the north coast of Spain (Figure 5). From May onwards, abundance increased gradually in the entire study area, reaching approximately 10,000 individuals per night in July and August in Germany (Figure 5). Abundance decreased slightly in September but increased again in October to approximately 10,000 *Culicoides* per night in Germany although the inter-annual variation also increased for October. After this, abundance decreased in November, with the highest abundance areas located in Germany. From December to February abundance was predicted to be very low (<10 specimens or null) (Figure 5). The coefficient of variation maps showed that the highest coefficient of variation between years was found in Spain, indicating that this area had the highest variation in the predictions between all the years (Figure 5).

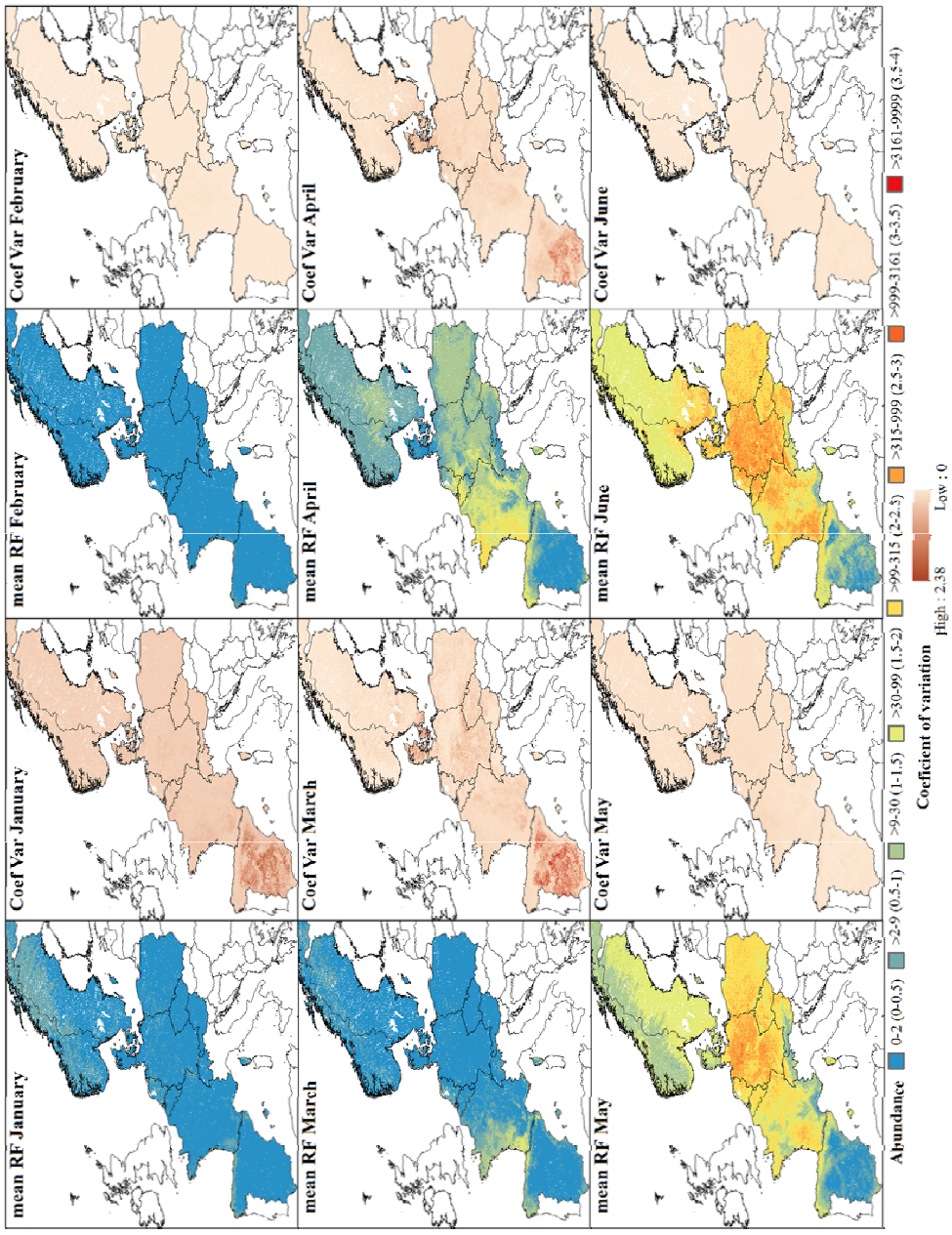


Figure 5. Abundance maps from January to June of the Obsoletus ensemble. The mean predictions were calculated per pixel using the seven prediction maps made for each year. Values are shown on a log10 scale. Coefficient of variation maps highlight the areas with the higher variation in the predictions for the seven year study period.

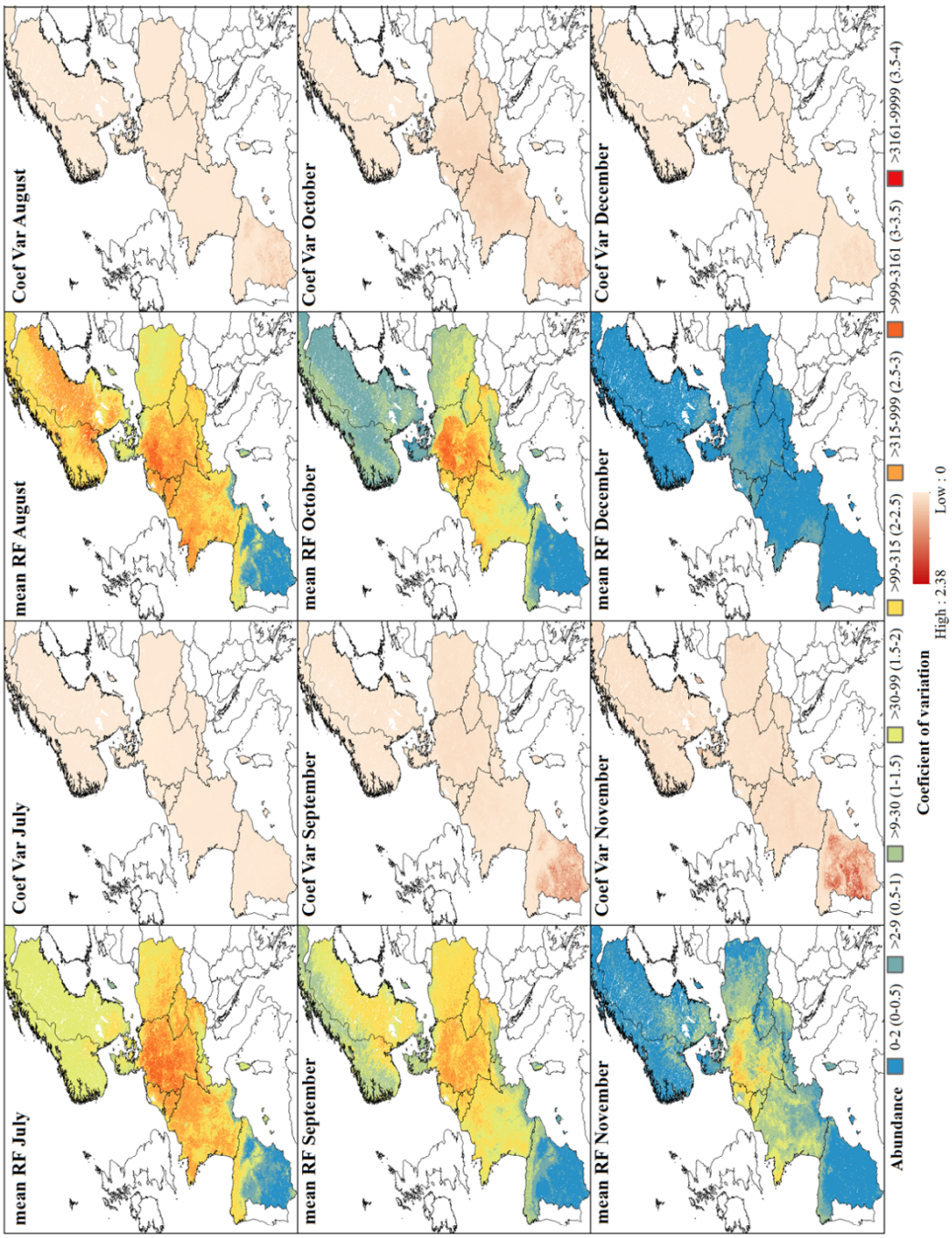


Figure 5. Abundance maps from July to December of the Obsoletus ensemble.

The predicted abundance for the Pulicaris ensemble, showed a similar seasonal pattern with a slow increase in abundance from April (Figure 6), until it peaked in May with a maximum prediction of approximately 2,800 individuals. March and April both showed a large inter-annual variation in the predictions starting with Germany, Poland and southern Scandinavia in March and spreading to most of the study area in April. During this month, the highest abundance was predicted in northern Germany, with a decreasing abundance towards western France and medium abundance towards Poland. This pattern was maintained until October, and in November abundance started to decrease gradually, with northern Germany having the highest abundance (Figure 6). In general, the Pulicaris ensemble showed a more easterly distribution (Germany, Poland and Scandinavia) compared to the Obsoletus ensemble and a much lower overall abundance.

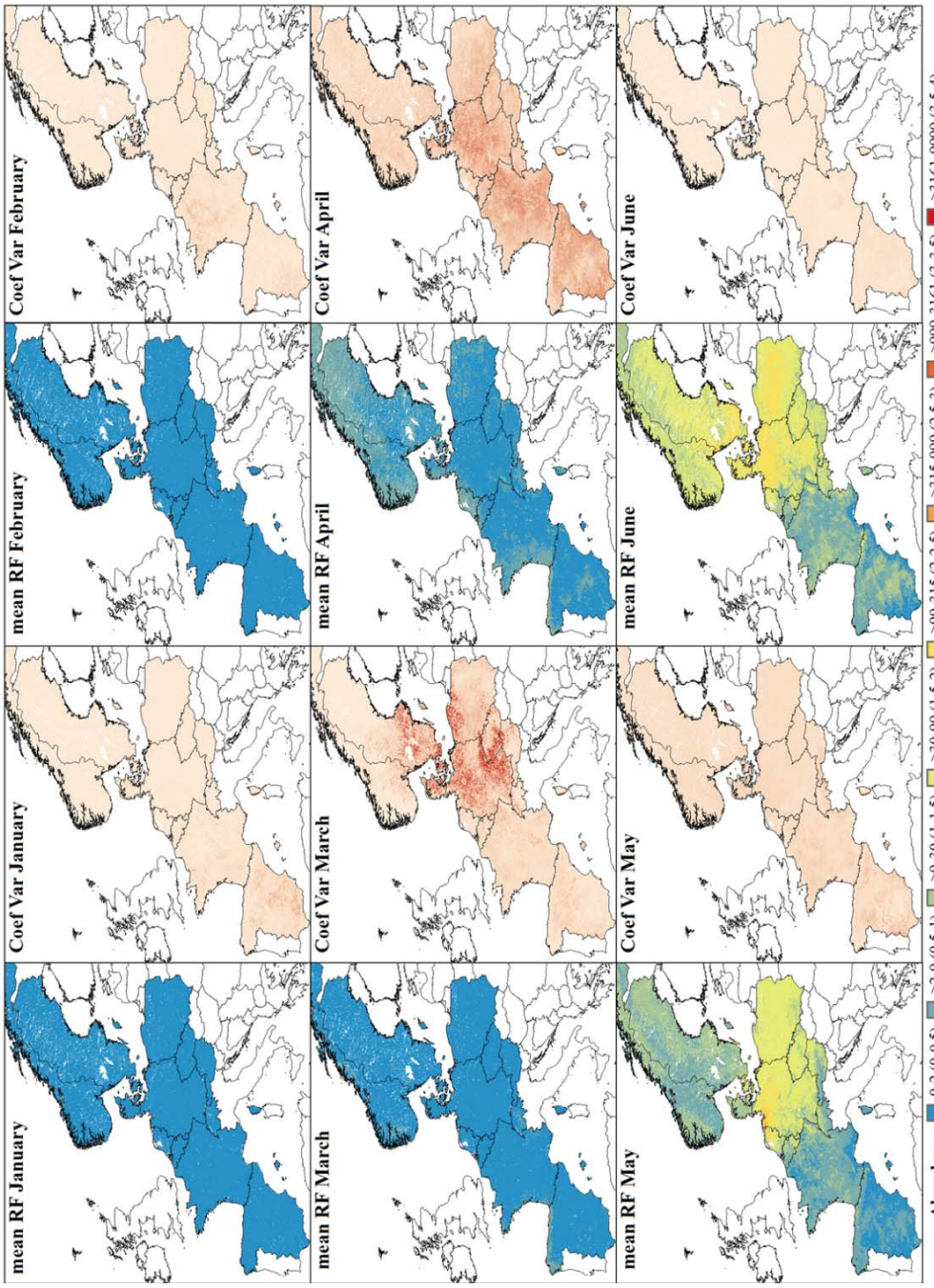


Figure 6. Abundance maps from January to June of the Pulicaris ensemble. The mean predictions were calculated per pixel using the seven prediction maps made for each year. Values are shown on a log10 scale. Coefficient of variation maps highlight the areas with the higher variation in the predictions for the seven year study period.

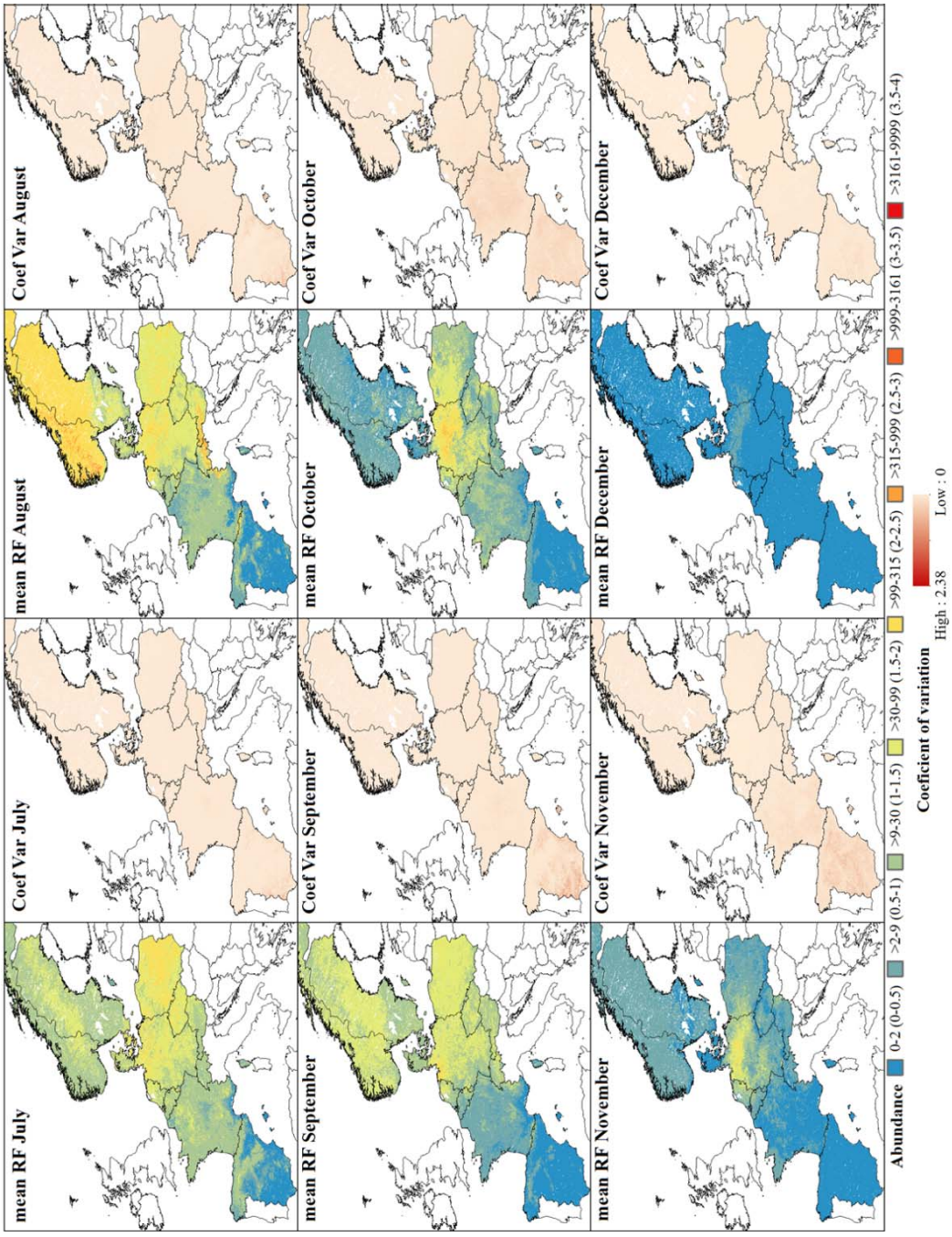


Figure 6. Abundance maps from July to December of the Pulicaris ensemble.

C. imicola was predicted to have very low abundance in January and February (<10 individuals), with the abundance increasing gradually in March, until it peaked in July and October in central Spain and on the coast of Corsica (Figures 7). *Culicoides imicola* was always predicted in the southwest and central areas of Spain and coastal areas in Corsica. A peak in the southern coast of Corsica was predicted during July (approximately 1,000 individuals) (Figure 7).

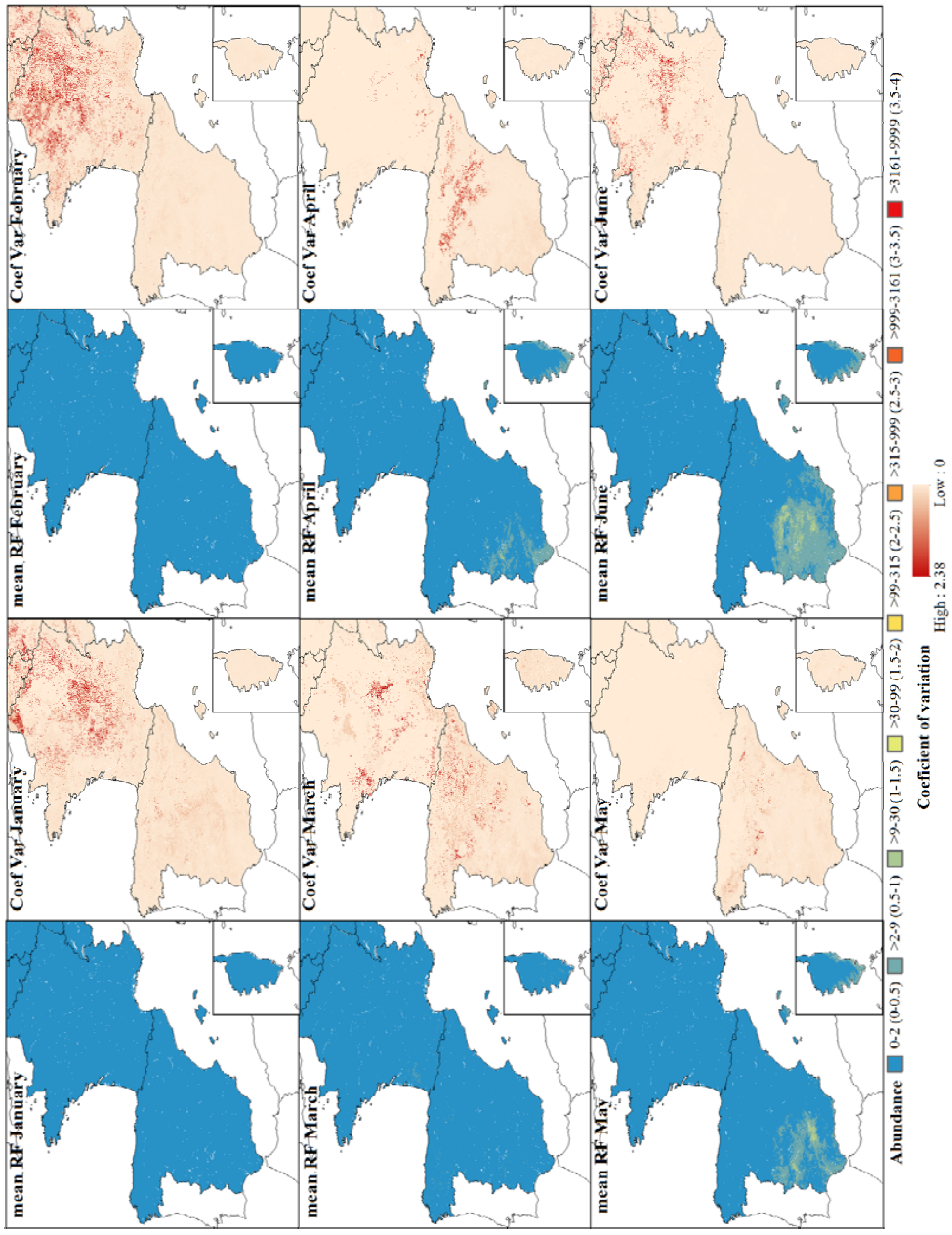


Figure 7. Abundance maps from January to June of *C. imicola* of the Iberian Peninsula and Corsica (displayed on the bottom right corner). The mean predictions were calculated per pixel using the seven prediction maps made for each year. Values are shown on a log10 scale. Coefficient of variation maps highlight the areas with the higher variation in the predictions for the seven year study period.

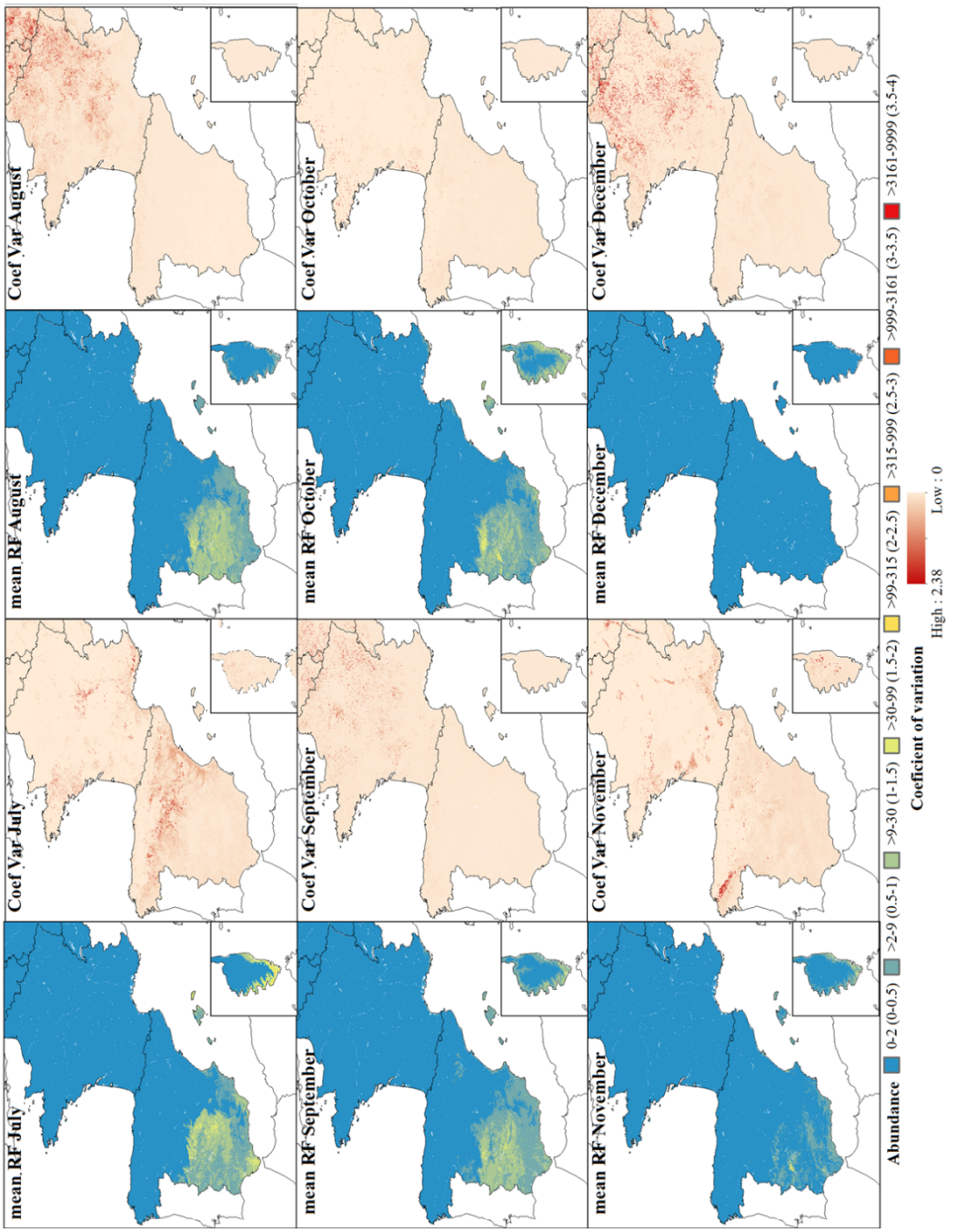


Figure 7. Abundance maps from July to December of *C. imicola* of the Iberian Peninsula and Corsica (displayed on the bottom right corner).

Variable importance

The five most important predictor variables identified for each month and for each *Culicoides* group are reported in Table 4

Month	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
January	**dLST A1 100.00** dLST DA 100.00 ***CLC_1k_P 100.00***	**EVI A0 99.29** BIO 1 97.00 ***NDVI A1 98.74***	**Altitude 94.14** rec_snow 96.35 ***EVI A0 97.40***	**year.2012 91.60** MIRA0 94.55 ***NDVI A3 95.75***	**dLST P2 83.34** nLST A1 91.93 ***dLST D1 94.60***
February	**dLST A1 100.00** EVI A1 100.00 ***MIR DA 100.00***	**BIO 2 99.94** dLST A0 96.31 ***BIO 1 98.36***	**EVI A0 97.78** MIR A2 91.27 ***nLST P3 94.50***	**MIR A2 95.94** dLST D2 91.10 ***BIO 14 93.88***	**MIR A0 91.06** MIR P1 90.59 ***EVI A1 93.60***
March	**dLST A1 100.00** nLST P3 100.00 ***MIR P2 100.00***	**year.2010 77.84** nLST A0 79.53 ***BIO 13 91.95***	**nLST A0 56.26** Goat 79.05 ***EVI P2 74.63***	**nLST A1 53.74** BIO 13 74.89 ***dLST P1 57.40***	**CLC 20 41.41** nLST DA 66.09 ***BIO 14 54.28***
April	**nLST A0 100.00** year.2011 100.00 ***BIO 1 100.00***	**nLST A2 97.67** year.2010 70.83 ***nLST A0 83.68***	**year.2011 90.51** nLST A0 49.93 ***BIO 15 73.06***	**year.2008 89.18** EVI A0 47.05 ***year.2008 72.51***	**BIO 1 78.46** year.2013 40.06 ***NDVI A0 70.26***
May	**year.2010 100.00** nLST P3 100.00 ***BIO 1 100.00***	**BIO 18 66.19** year.2010 93.51 ***BIO 15 91.95***	**BIO 8 48.85** BIO 8 63.29 ***nLST A1 91.21***	**BIO 2 44.25** BIO 1 55.73 ***BIO 14 91.03***	**Altitude 41.94** MIR A0 51.14 ***nLST A0 86.77***
June	**BIO 18 100.00** BIO 1 100.00 ***nLST A0 100.00***	**dLST P3 50.59** nLST P2 61.17 ***BIO 15 99.11***	**nLST P1 46.58** nLST P3 56.51 ***BIO 18 94.78***	**Chicken 46.50** NDVI A1 54.42 ***BIO 1 91.18***	**MIR P2 45.15** year.2008 54.06 ***BIO 14 90.86***
July	**BIO 18 100.00** nLST P3 100.00 ***BIO 14 100.00***	**BIO 14 98.33** BIO 1 81.05 ***nLST A1 50.93***	**BIO 2 97.42** dLST P3 73.82 ***BIO 13 44.93***	**MIR DA 84.94** BIO 14 73.02 ***BIO 1 40.98***	**nLST P1 82.09** dLST A0 67.65 ***CLC 21 40.42***
August	**dLST A0 100.00** dLST A0 100.00 ***BIO 14 100.00***	**BIO 18 91.33** BIO 1 91.09 ***BIO 2 95.02***	**BIO 14 90.80** year.2012 60.46 ***BIO 7 93.42***	**dLST P3 76.40** dLST P3 58.43 ***BIO 15 92.89***	**BIO 3 74.36** MIR P3 57.82 ***dLST A0 92.69***
September	**BIO 18 100.00** EVI P2 100.00 ***BIO 7 100.00***	**year.2012 86.72** BIO 1 99.29 ***nLST A0 98.87***	**nLST P1 80.28** nLST A2 86.95 ***BIO 13 97.98***	**nLST A2 57.82** dLST A0 84.97 ***MIR DA 95.11***	**MIR P2 48.07** nLST A0 84.67 ***BIO 2 94.76***
October	**year.2012 100.00** nLST A2 100.00 ***BIO 14 100.00***	**BIO 3 37.34** year.2012 99.18 ***BIO 1 83.92***	**BIO 18 32.36** BIO 8 56.73 ***nLST A0 74.26***	**nLST A2 24.67** cr_1_18 53.30 ***BIO 15 67.94***	**BIO 14 23.31** year.2008 47.65 ***BIO 18 67.85***
November	**nLST A2 100.00** nLST P3 100.00 ***BIO 14 100.00***	**year.2012 65.02** BIO 8 84.09 ***BIO 1 98.49***	**EVI A0 62.95** nLST A2 79.19 ***CLC 18 82.93***	**year.2011 59.88** dLST P2 67.91 ***MIR DA 69.08***	**nLST P3 56.11** Altitude 65.13 ***dLST P2 65.40***
December	**Altitude 100.00** BIO 2 100.00 ***BIO 2 100.00***	**NDVI A0 97.44** EVI P2 95.35 ***dLST A1 85.72***	**dLST A1 97.30** BIO 1 93.31 ***BIO 18 81.66***	**EVI A2 92.83** NDVI A2 90.91 ***EVI D1 77.56***	**EVI D2 92.24** BIO 13 90.14 ***nLST A1 76.39***

Table 4. The five most important variables given by RF models for each month. In each cell the Obsoletus ensemble is shown as 1st row (in bold capital letters), the Pulicaris ensemble is shown as 2nd row (in capital letters) and *C. imicola.*is shown as 3rd row (in Bold, capital, italic letters): The number indicates the importance rank of the variable. The top most important variables ("variable 1" column) have a value of 100.

In general the most important variables for the Obsoletus ensemble were related to temperature (dLST) and to precipitation (BIO 18). For the months of May and October, year 2010 and year 2012 were the most important variables respectively.

For the Pulicaris ensemble, the most important variables were related to temperature and vegetation index. For *C. imicola* the most important variables were related to pastures, precipitation and temperature.

Interpolation based abundance maps

Interpolation of the monthly average abundance was performed for each month, resulting in 12 average abundance maps. Interpolation was only carried out for the area between the most northern, southern, western and eastern farm in the training dataset each month.

RF models were run using the same training set as in the interpolation, in order to directly compare RF results to the interpolation results. Both models were validated against the same test data set and comparisons of the interpolation results with the RF models are shown in table 5.

	Obsoletus		**Pulicaris**		***C. imicola***	
Month	**nRMSE RF**	**nRMSE Interpolation**	**nRMSE RF**	**nRMSE Interpolation**	**nRMSE RF**	**nRMSE Interpolation**
January	1.73	2.25	2	3.92	6.03	6.98
February	2.27	2.45	4.33	3.83	5.73	5.22
March	0.98	0.98	2.56	3.29	6.04	11.2
April	0.48	0.55	0.68	0.69	2.62	2.51
May	0.53	0.58	0.9	0.98	3.31	3.28
June	0.47	0.47	0.73	0.73	2.30	2.58
July	0.47	0.5	0.68	0.71	2.27	2.74
August	0.64	0.65	1.13	1.3	3.20	3.86
September	0.53	0.55	0.78	0.78	1.95	2.23
October	0.42	0.49	0.68	0.72	2.41	2.30
November	0.68	0.7	1	1	2.28	2.68
December	1.46	1.48	3.11	3.18	10.86	10.2
Total mean	0.88	0.97	1.54	1.76	4.08	4.65

Table 5. nRMSE values for the RF models and the interpolation for the months of January to December. RF and interpolation were performed using the average abundance. The mean for all the months for each method are shown in the last row.

Comparing the mean nRMSE for the three *Culicoides* taxa/species, the RF model performed slightly better than the interpolation (table 5). Considering the nRMSE for the Obsoletus ensemble alone, interpolation seemed to perform better than RF models for February while for *C. imicola*, interpolation seemed to perform better for the months of February, April, May, and October (table 5). In general, the scatterplots for the predicted and observed values for both the interpolation and the RF models were quite similar (Additional file1: Figures S1, S2, S3), however, the interpolation models predicted a larger abundance dispersal and range compared to the RF models.The range predicted by the interpolation method was closer to the observed range than the more limited range predicted by the RF method, eventhough the interpolation predictions were not more precise than RF predictions (i.e. they were not closer to the best fit line) (Additional file1: Figures S1, S2, S3). When observing the abundance maps obtained from both methods, the RF maps seems to be smoother compared to the interpolation maps. This is because the interpolation maps are showing higher predicted values in the sourroundings of the farms used for training. However, when zooming in on the maps, it can be seen that the interpolation models resulted in a smooth transition from farm to farm while the predictions from the environment driven RF actually vary pixel by pixel (Additional file 1: Figure S7)

Discussion

We modelled the abundance of the Obsoletus and Pulicaris ensembles and of *C. imicola* using the machine learning technique Random Forest (RF) and predicted the vector abundance at a continental scale using entomological data obtained from national monitoring programs in nine European countries. We used abundance catch data from 31,429 *Culicoides* traps covering the years 2007 to 2013. The predicted abundance maps presented here constitute the first *Culicoides* abundance models made for a large part of Europe and were based on the largest entomological dataset generated to date for *Culicoides*. Besides showing the major geographic abundance patterns and giving an

insight into the seasonal dynamics on a monthly scale, the prediction maps presented here are also useful for risk assessment and for the development of R_0 models for both established or future emerging *Culicoides* borne diseases.

In this study, we used RF technique to predict vector abundance. RF performance for predicting the abundance varied with season. In general, the error, measured as the nRMSE was higher during the winter months compared to the warmer months, possibly due to lower amount of farms sampled during the winter. The less than fair performance obtained from our RF models for some months, may be explained by limitations of the RF algorithm, the lack of important predictors or by limitations caused by using a data set merged from different sources. In the RF algorithm, predictions for extreme observations (low or high values) are computed by averaging the training data set outcomes in the terminal nodes and therefore, large values will necessarily be underestimated and low values overestimated [59]. Another reason for low performance may be that the remote sensing predictors used here were not the key drivers or not the only key drivers of *Culicoides* abundance on European farms and maybe instead landscape conditions at a finer scale, such as farm practice, management and micro habitats are important drivers of vector abundance at the farms. For instance, the presence of dung heaps has been associated with the emergence of adult *C. obsoletus*, while cow pats were associated with *C. dewulfi* and *C. chiopterus* in Germany [61]. These are variables that cannot be detected by remote sensing and therefore have to be extracted from industry owned farm databases or be measured manually when trapping for *Culicoides*. Enhanced predictions will be only possible for farms as long as they have managements practices recorded in available databases. Low performance of our models could also be due to the resolution of the predictors. $1km^2$ may not be the optimal resolution for capturing certain landscape features at farm scale, like moist soil conditions determining the presence of small breeding sites. For example, *C. imicola* oviposits on mud or semi- moist areas, at the margin of ponds or close to leaking irrigation pipes [24,62]. The use of high resolution satellite images may help to identify these areas and improve the power of future model predictions.

Another source of error comes from the vector data used: the mean observed abundance per month showed large variation among the seven different years sampled (data not shown). Large inter-annual variation poses a problem for a model to exactly predict the observed values, as any value within the observed range can be expected. Here, we attempted to correct for inter-annual variation by adding "year" as a categorical variable. We did this to train the RF model with information regarding the year in which the collections were made, resulting in predictions for that particular year. Adding year as categorical variable resulted in a model only being able to predict observations belonging to the period sampled and thus the models cannot be used to predict to future years. We wanted to generate a prediction map that could be used for decision making under possible future outbreaks scenarios; therefore we took the average of all the prediction maps, assuming the average of the past seven years is the best prediction for the future. This map can be interpreted as a general overview of what to expect regarding average monthly *Culicoides* abundance.

Lastly, when working with spatial data collected in different years, comparing data from different regions may be problematic as it will not be clear if any difference found is reflecting a real abundance difference in the spatial domain or if it is a product of special conditions occurring during the years of sampling (e.g. extreme temperatures during the hot season).

Random Forest is a non-spatial and predictor based modelling approach. We compared the results of RF with the results obtained from a purely spatial interpolation method (Inversed Distance Weighting, IDW) that does not require any predictor variable. Creating maps using interpolation is a simple method and predictions to un-sampled regions are made by calculating the average of the values measured at neighbouring locations assuming that high and low abundance cluster spatially, thus that abundance on one farm can be predicted by abundances on close by farms. IDW performed on average only slightly worse than the RF models using the exact same training and test sets, probably because the average abundance at the farms is not sufficiently spatially auto correlated; otherwise,

interpolation methods would have been able to predict the abundance more precisely than RF.

Differences between both methods were observed according to the scale: at a local scale, the interpolation method resulted in generally smooth surfaced maps for neighbouring farms (additional file 1, S7). This makes sense as interpolation is a method fitting a function using the surrounding points, and therefore, it is expected to produce a smooth surface between two neighbouring points. On the contrary, the RF produced more spatially variable abundance maps at a local scale, suggesting that predictor variables with large variation at the local scale, such as land cover, are important drivers in the RF models. The RF models however, did not perform dramatically better than simple interpolation methods, suggesting that the available land cover classes used to train the model only have a limited predictive value of *Culicoides* abundance. The lack of importance of land cover in predicting vector abundance is also supported by the RF decision trees mainly selecting climate variables rather than land cover variables as predictors. At a continental scale, the interpolated maps showed a few marked hotspots and seemed to be more variable than the RF maps. This is not reflecting more precise predictions, but a larger range of predicted abundance, probably also with a low predictive power. Conversely, the RF models produced a smoother spatial abundance, as a result of a more narrow range in predictions, mainly due to the fact that RF tends to overestimate low values and underestimate high values. Since the interpolation predictions were fair, we conclude that a large part of the variation in the abundance across Europe is explained by drivers with a gradual change. Temperature has a fairly smooth transition from southern to northern Europe and therefore temperature related variables are likely to be the underlying variables driving the continental scale abundance distribution.

For some months, there was a fair relation between the predicted and observed abundance. We calculated the nRMSE using all the farms belonging to the test sets, without considering the country – resulting in a fairly low nRMSE values. However, separating the predicted abundance by country showed that the models were not able to predict the observed abundance within the individual countries very well. Therefore, our models were

able to predict average abundance in large regions in Europe and distinguish Spain from Germany and from the Scandinavian countries based on mainly climatic variables but they were not good at predicting variation in abundance within regions in a country and much less at farm level were the climate is identical but the variation in abundance is driven by non-climatic variables.

Land cover has been shown to have an effect on the occurrence of *C. impunctatus* and *C. punctatus* in Scotland [45], of Obsoletus complex in Italy [63,64], and for *Culicoides* spp in Germany [65] but not for Palearctic *Culicoides* species in Denmark [66]. In this study none of the land cover type considered, appeared as important predictors. This may partly be explained by the fact that the percentage of land cover for a given farm was assigned extracting the value from the pixel to which the farm belonged to, without considering the position of the farm inside a pixel. Thus, farms located in the corner of the pixel might be more related to the neighbouring values than its own pixel value. It is important to note that all the vector data used here are from farms and therefore inherently have a very limited range of land covers. Had vector data been collected at random points in Europe and not only at farms then land cover variables would likely have had a much larger effect as the *Culicoides* vectors in this study are highly associated with farms.

For the Obsoletus ensemble, the most important predictors varied throughout the year, however they were all related to temperature (dLST, nLST) and precipitation (BIO 18). Obsoletus ensemble species have a Palearctic distribution and they are widely distributed in central and northern Europe, with low abundance or absence in central and southern Spain. The Obsoletus ensemble distribution coincides with the humid oceanic climates, characterized by warm summers, and with the temperate and humid continental climate [67]. Obsoletus has also been reported to prefer colder environments where rainfall is regular throughout the year [15]. The model identified the highest abundance areas for the Obsoletus ensemble to be in Germany followed by France. The southern coast of Norway presented the highest catch of all the observations in the data set [46]. The models did not predict these high abundances for Norway, maybe because the extreme catch observed in the data set is not usual for the area due to the environmental conditions there, or it is

possible that the Obsoletus ensemble is established there in high abundance, but because RF underestimate at higher values, this area ended up with lower predictions than the observed data.

Versteit et al presented an abundance map for the Obsoletus ensemble in Europe made by Balenghien and Wint [39]. The spatial pattern shown in our maps is relatively similar, differing in the high abundance located in Germany in our prediction maps, where they showed the highest abundance to be located in western France. Our maps also differed in respect to the high abundance area Versteit et al. showed in the Carpathian Mountains in Eastern Europe. Our maps were unable to distinguish these regions as high abundance areas. Their map was made with abundance data collected in Iceland, UK, Portugal, Denmark, Norway, and Finland and the mean maximum catch was predicted. In the same report, the authors show another abundance map made by Withenshaw et al [38], in which they modelled the maximum catch from collections made in the UK and Spain. For the Obsoletus ensemble, their map showed higher abundances at higher latitudes, decreasing as latitude decreased. Their results might be reflecting high Obsoletus abundances in UK, and the model extrapolates high abundances to environmentally similar areas in Europe.

For the Pulicaris ensemble, the most important variables for the months where the model performed fairly well were related to temperature (BIO 1, BIO2, nLST, dLST). As the Obsoletus ensemble, the Pulicaris ensemble has been found to occur in cool and wet climates (down to 7 °C annual mean and up to 700 mm of rainfall). Our maps showed that the Pulicaris ensemble was widely distributed in Europe with the highest abundance occurring in northern Germany, where Pulicaris abundance was reported to be extremely high in some locations [68], and with high abundance occurring in Poland, in accordance with other studies [46].

The RF models for *C. imicola* had the lowest performance compared to both ensembles. Nevertheless, our resulting maps displayed *C. imicola* abundance similar to previous studies where *C. imicola* abundance was modelled in Spain [35]. Our models were able to recognize environmental factors at a regional scale that allowed us to estimate quite

accurately the distribution of *C. imicola*, as our maps are comparable to maps presented by other studies in Spain [30,31,34].

The most important variables identified by our models were related to temperature (BIO1, nLST) and precipitation (BIO 14). Annual mean temperature has been reported to be the main driver of *C. imicola* in Europe. The species occurs where temperatures are high on average and stable during the year [33,69]. Precipitation has also been known to affect the species, as they mostly occur where rainfall is below 700 mm annually. *C. imicola* pupa cannot float and could drown under abundant rainfall conditions [31,69,70].

R_0 models for vector borne diseases depend, among other parameters, on vector abundance (or vector density). Vector density is used to calculate the vector-to-host-ratio (denoted by m) as m= N/H, where N is the vector density and H the host density [16]. This parameter appears in R_0 equations as a factor and therefore, a higher host density (at a constant vector abundance) leads to a lower R_0, while a higher vector density (at a constant host abundance) leads to a higher R_0. Some authors have calculated R_0 for *Culicoides*-borne BTV in Netherlands [17], Austria [71] and Europe [72] and for African horse sickness [73]. In this work, we calculated biting midge abundance in Europe at a large regional scale and therefore these results, in some extent, are useful to calculate R_0 values for any *Culicoides-borne* disease in Europe. Until date, no consensus has been made on how to calculate vector-to-host-ratio from collections made by light traps. Some authors considered collections in light trap catches as 1 % of the total vector population at a farm, and therefore they multiplied trap catches by 100 [17,71] and divided it by the number of hosts at the farm. In this approach, the calculation of the vector to host ratio is dependent of the number of hosts present on a farm. In turn, Guis et al [72] directly used the trap catch abundance as the vector-to-host ratio as they considered the light trap to act as a host and to attract the same number of midges as a host would attract. In the later approach, calculation of the vector to host ratios is independent on the number of hosts.

Conclusions

The model performance varied with the species or *Culicoides* ensembles, with the Obsoletus ensemble models having the highest performance and *C. imicola* models the lowest performance. Model performance also varied with season, with the lowest performance found during the winter months. Our RF models were able to distinguish average abundance between different regions within nine European countries, but gave poor predictions of the relatively large observed variation in abundance at a farm scale, potentially due to model limitations, predictor resolution, or lack of important predictor variables. With the high number of trap data used, we could predict *Culicoides* abundance at farm level with nearly the same average precision using a simple interpolation approach as when using an advanced environmental predictor-driven modelling approach. For the Obsoletus ensemble, model predictions were fair; indicating that the maps produced here can be used as input for more general modelling approaches, such as R_0 models for *Culicoides*-borne disease risk assessment.

References

1. Carpenter S, Wilson A, Mellor PS. Culicoides and the emergence of bluetongue virus in northern Europe. Trends Microbiol. [Internet]. 2009;17:172–8. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0966842X09000407

2. Pinior B, Brugger K, Kofer J, Schwermer H, Stockreiter S, Loitsch A, et al. Economic comparison of the monitoring programmes for bluetongue vectors in Austria and Switzerland. Vet. Rec. [Internet]. 2015;176:464–464. Available from: http://veterinaryrecord.bmj.com/cgi/doi/10.1136/vr.102979

3. Rushton J, Lyons N. Economic impact of Bluetongue: a review of the effects on production. Vet. Ital. [Internet]. 2015;51:401–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26741252

4. Mellor PS, Carpenter S, Harrup L, Baylis M, Mertens PPC. Bluetongue in Europe and the Mediterranean Basin: History of occurrence prior to 2006. Prev. Vet. Med. [Internet]. 2008;87:4–20. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0167587708001189

5. Toussaint J-F, Sailleau C, Mast J, Houdart P, Czaplicki G, Demeestere L, et al. Bluetongue in Belgium, 2006. Emerg. Infect. Dis. [Internet]. 2007;13:614–6. Available from: http://wwwnc.cdc.gov/eid/article/13/4/06-1136_article.htm

6. Saegerman C, Berkvens D, Mellor PS. Bluetongue epidemiology in the European Union. Emerg. Infect. Dis. 2008;14:539–44.

7. Sperlova A, Zendulkova D. Bluetongue: A review. Vet. Med. (Praha). 2011;56:430–52.

8. Wilson a, Carpenter S, Gloster J, Mellor P. Re-emergence of bluetongue in northern Europe in 2007. Vet. Rec. [Internet]. 2007;161:487–9. Available from: http://veterinaryrecord.bmj.com/cgi/doi/10.1136/vr.161.14.487

9. EFSA Panel on Animal Health and Welfare. Bluetongue: control, surveillance and safe movement of animals. EFSA J. [Internet]. 2017;15. Available from: http://doi.wiley.com/10.2903/j.efsa.2017.4698

10. Meiswinkel R, Baldet T, de Deken R, Takken W, Delécolle J-C, Mellor PS. The 2006 outbreak of bluetongue in northern Europe—The entomological perspective. Prev. Vet. Med. [Internet]. 2008;87:55–63. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0167587708001220

11. Mehlhorn H, Walldorf V, Klimpel S, Jahn B, Jaeger F, Eschweiler J, et al. First

occurrence of Culicoides obsoletus-transmitted Bluetongue virus epidemic in Central Europe. Parasitol. Res. [Internet]. 2007;101:219–28. Available from: http://link.springer.com/10.1007/s00436-007-0519-6

12. Hoffmann B, Bauer B, Bauer C, Bätza HJ, Beer M, Clausen PH, et al. Monitoring of putative vectors of bluetongue virus serotype 8, Germany. Emerg. Infect. Dis. 2009;15:1481–4.

13. Meiswinkel R, van Rijn P, Leijs P, Goffredo M. Potential new Culicoides vector of bluetongue virus in northern Europe. Vet. Rec. [Internet]. 2007;161:564–5. Available from: http://veterinaryrecord.bmj.com/cgi/doi/10.1136/vr.161.16.564

14. Dijkstra E, van der Ven IJK, Meiswinkel R, Holzel DR, van Rijn PA, Meiswinkel R. Culicoides chiopterus as a potential vector of bluetongue virus in Europe. Vet. Rec. 2008;162:422–422.

15. Venail R, Balenghien T, Guis H, Tran A, Setier-Rio M-L, Delécolle J-C, et al. Assessing Diversity and Abundance of Vector Populations at a National Scale: Example of Culicoides Surveillance in France After Bluetongue Virus Emergence. In: Mehlhorn H, editor. Arthropods As Vectors Emerg. Dis. [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 77–102. Available from: http://link.springer.com/10.1007/978-3-642-28842-5_4

16. Gubbins S, Carpenter S, Baylis M, Wood JL., Mellor PS. Assessing the risk of bluetongue to UK livestock: uncertainty and sensitivity analyses of a temperature-dependent model for the basic reproduction number. J. R. Soc. Interface [Internet]. 2008;5:363–71. Available from: http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2007.1110

17. Hartemink NA, Purse BV, Meiswinkel R, Brown HE, de Koeijer A, Elbers ARW, et al. Mapping the basic reproduction number (R0) for vector-borne diseases: A case study on bluetongue virus. Epidemics [Internet]. 2009;1:153–61. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1755436509000309

18. Elith J, Leathwick JR. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu. Rev. Ecol. Evol. Syst. 2009;40:677–97.

19. Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. Global Environmental Data for Mapping Infectious Disease Distribution. Adv. Parasitol. 2006;62:37–77.

20. Wittman EJ, Baylis M. Climate Change: Effects on Culicoides -Transmitted Viruses and Implications for the UK. Vet. J. [Internet]. 2000;160:107–17. Available from:

http://linkinghub.elsevier.com/retrieve/pii/S1090023300904702

21. Mullens BA, Gerry AC, Lysyk TJ, Schmidtmann ET. Environmental effects on vector competence and virogenesis of bluetongue virus in Culicoides: interpreting laboratory data in a field context. Vet. Ital. [Internet]. 2004;40:160–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20419655

22. Mellor PS, Boorman J, Baylis M. C ULICOIDES B ITING M IDGES : Their Role as Arbovirus Vectors. 2000;307–40.

23. Kettle DS. The Bionomics and Control of Culicoides and Leptoconops (Diptera, Ceratopogonidae = Heleidae). Annu. Rev. Entomol. [Internet]. 1962;7:401–18. Available from: <Go to ISI>://19580500007

24. Mellor PS, Prrzous G. Observations on breeding sites and light-trap collections of Culicoides during an outbreak of bluetongue in Cyprus. Bull. Entomol. Res. 1979;69:229–34.

25. Purse B V, Rogers D. Bluetongue virus and climate change. In: Mellor P, Baylis M, Mertens PPC, editors. Bluetongue. Academic Press; 2008.

26. Kalluri S, Gilruth P, Rogers D, Szczur M. Surveillance of Arthropod Vector-Borne Infectious Diseases Using Remote Sensing Techniques: A Review. PLoS Pathog. [Internet]. 2007;3:e116. Available from: http://dx.plos.org/10.1371/journal.ppat.0030116

27. van Doninck J, De Baets B, Peters J, Hendrickx G, Ducheyne E, Verhoest NEC. Modelling the spatial distribution of Culicoides imicola: Climatic versus remote sensing data. Remote Sens. 2014;6:6604–19.

28. Conte A, Giovannini A, Savini L, Goffredo M, Calistri P, Meiswinkel R. The effect of climate on the presence of Culicoides imicola in Italy. J. Vet. Med. Ser. B. 2003;50:139–47.

29. Purse B V, Tatem a J, Caracappa S, Rogers DJ, Mellor PS, Baylis M, et al. Modelling the distributions of Culicoides bluetongue virus vectors in Sicily in relation to satellite-derived climate variables. Med. Vet. Entomol. [Internet]. 2004;18:90–101. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.0269-283X.2004.00492.x/pdf

30. Peters J, De Baets B, Van Doninck J, Calvete C, Lucientes J, De Clercq EM, et al. Absence reduction in entomological surveillance data to improve niche-based distribution models for Culicoides imicola. Prev. Vet. Med. [Internet]. Elsevier B.V.; 2011;100:15–28. Available from: http://dx.doi.org/10.1016/j.prevetmed.2011.03.004

31. Calvete C, Estrada R, Miranda MA, Borrás D, Calvo JH, Lucientes J. Modelling the distributions and spatial coincidence of bluetongue vectors Culicoides imicola and the Culicoides obsoletus group throughout the Iberian peninsula. Med. Vet. Entomol. [Internet]. 2008;22:124–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18498611

32. Tatem AJ, Baylis M, Mellor PS, Purse B V., Capela R, Pena I, et al. Prediction of bluetongue vector distribution in Europe and north Africa using satellite imagery. Vet. Microbiol. [Internet]. 2003;97:13–29. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0378113503002979

33. Purse B, Mccormick BJJ, Mellor PS, Baylis M, Boorman JPT, Borras D, et al. Incriminating bluetongue virus vectors with climate envelope models. J. Appl. Ecol. [Internet]. 2007;44:1231–42. Available from: http://doi.wiley.com/10.1111/j.1365-2664.2007.01342.x

34. Ducheyne E, Miranda Chueca MA, Lucientes J, Calvete C, Estrada R, Boender G, et al. Abundance modelling of invasive and indigenous Culicoides species in Spain. Geospat. Health [Internet]. 2013;8:241–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24258899

35. Acevedo P, Ruiz-Fons F, Estrada R, Márquez AL, Miranda MA, Gortázar C, et al. A Broad Assessment of Factors Determining Culicoides imicola Abundance: Modelling the Present and Forecasting Its Future in Climate Change Scenarios. Cornell SJ, editor. PLoS One [Internet]. 2010;5:e14236. Available from: http://dx.plos.org/10.1371/journal.pone.0014236

36. Rigot T, Conte A, Goffredo M, Ducheyne E, Hendrickx G, Gilbert M. Predicting the spatio-temporal distribution of Culicoides imicola in Sardinia using a discrete-time population model. Parasit. Vectors [Internet]. Parasites & Vectors; 2012;5:270. Available from: Parasites & Vectors

37. Baylis M, Mellor PS, Wittmann EJ, Rogers DJ. Prediction of areas around the Mediterranean at risk of bluetongue by modelling the distribution of its vector using satellite imaging. Vet. Rec. 2001;149:639–43.

38. Withenshaw S, Searle KR, Butler A, Allepuz A, Barber J, Carpenter S, et al. Modelling the roles of climate, landscape and biotic factors in the distribution and maximum trap catch of Culicoides vectors across Europe [Internet]. Available from: http://geri2015.edenext.eu/content/download/4211/31473/version/1/file/P+4.9.pdf

39. Versteirt V, Balenghien T, Tack W, Wint W. A first estimation of Culicoides imicola

and Culicoides obsoletus/Culicoides scoticus seasonality and abundance in Europe. EFSA Support. Publ. [Internet]. 2017;14. Available from: http://doi.wiley.com/10.2903/sp.efsa.2017.EN-1182

40. Kuhn M, Johnson K. Applied Predictive Modeling [Internet]. New York, NY: Springer New York; 2013. Available from: http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/ref=pd_bxgy_b_img_z

41. Pili E, Ciuccé S, Culurgioni J, Figus V, Pinna G, Marchi A. Distribution and abundance of bluetongue vectors in Sardinia: Comparison of field data with prediction maps. J. Vet. Med. Ser. B Infect. Dis. Vet. Public Heal. 2006;53:312–6.

42. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random Forests for Classification in Ecology. Ecology. 2007;88:2783–92.

43. Cianci D, Hartemink N, Ibáñez-Justicia A. Modelling the potential spatial distribution of mosquito species using three different techniques. Int. J. Health Geogr. [Internet]. 2015;14:10. Available from: http://www.ij-healthgeographics.com/content/14/1/10

44. Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. Global Environmental Data for Mapping Infectious Disease Distribution. 2006. p. 37–77. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0065308X05620027

45. PURSE B V., FALCONER D, SULLIVAN MJ, CARPENTER S, MELLOR PS, PIERTNEY SB, et al. Impacts of climate, host and landscape factors on Culicoides species in Scotland. Med. Vet. Entomol. [Internet]. 2012;26:168–77. Available from: http://doi.wiley.com/10.1111/j.1365-2915.2011.00991.x

46. Cuéllar AC, Kjær LJ, Kirkeby C, Skovgard H, Nielsen SA, Stockmarr A, et al. Spatial and temporal variation in the abundance of Culicoides biting midges (Diptera: Ceratopogonidae) in nine European countries. Parasit. Vectors [Internet]. Parasites & Vectors; 2018;11:112. Available from: https://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-018-2706-y

47. Cuéllar AC, Kjær LJ, Skovgard H, Nielsen SA, Stockmarr A, Andersson G, et al. Monthly variation in the probability of presence of adult Culicoides populations in nine European countries and the implications for targeted surveillance. Parasit. Vectors. 2018;In Press.

48. Scharlemann JPW, Benz D, Hay SI, Purse B V., Tatem AJ, Wint GRW, et al. Global Data for Ecology and Epidemiology: A Novel Algorithm for Temporal Fourier Processing MODIS Data. Gething P, editor. PLoS One [Internet]. 2008;3:e1408. Available from:

http://dx.plos.org/10.1371/journal.pone.0001408

49. EDENext. Biology and control of vector-borne infections in Europe [Internet]. 2014. Available from: https://www.edenext.eu/

50. Hijmans RJ. Worldclim - Global Climate Data. Free climate data for ecological modeling and GIS [Internet]. 2017 [cited 2015 May 21]. Available from: http://www.worldclim.org/node/1

51. Robinson TP, Wint GRW, Conchedda G, Van Boeckel TP, Ercoli V, Palamara E, et al. Mapping the Global Distribution of Livestock. Baylis M, editor. PLoS One [Internet]. 2014;9. Available from: http://dx.plos.org/10.1371/journal.pone.0096084

52. European Environment Agency. Corine Land Cover. 2018; Available from: https://www.eea.europa.eu/data-and-maps/data/clc-2006-raster-3

53. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 2013.

54. Hijmans RJ. raster: Geographic Data Analysis and Modeling. R package version 2.5-8. [Internet]. 2016. Available from: https://cran.r-project.org/package=raster

55. Breiman L. Random Forests. Mach. Learn. [Internet]. 2001;45:5–32. Available from: http://link.springer.com/10.1023/A:1010933404324

56. Evans JS, Murphy MA, Holden ZA, Cushman SA. Modeling Species Distribution and Change Using Random Forest. 2011;139–59.

57. Peters J, Waegeman W, Van J, Ducheyne E, Calvete C, Lucientes J, et al. Predicting spatio-temporal Culicoides imicola distributions in Spain based on environmental habitat characteristics and species dispersal. Ecol. Inform. [Internet]. Elsevier B.V.; 2014;22:69–80. Available from: http://dx.doi.org/10.1016/j.ecoinf.2014.05.006

58. Ducheyne E, Charlier J, Vercruysse J, Rinaldi L, Biggeri A, Demeler J, et al. Modelling the spatial distribution of Fasciola hepatica in dairy cattle in Europe. Geospat. Health [Internet]. 2015;9:261–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25826307

59. Kuhn M. Building Predictive Models in R Using the caret Package. J. Stat. Softw. [Internet]. 2008;28:159–60. Available from: http://www.jstatsoft.org/v28/i05/

60. Liaw A, Wiener M. Classification and Regression by randomForest. R news. 2002;2/3:18–22.

61. Steinke S, Lühken R, Balczun C, Kiel E. Emergence of Culicoides obsoletus group species from farm-associated habitats in Germany. Med. Vet. Entomol. [Internet]. 2016;30:174–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26744290

62. Foxi C, Pinna M, Monteys VSI. An Updated Checklist of the Culicoides Latreille (Diptera : Ceratopogonidae) of Sardinia (Italy), and Seasonality in Proven and Potential Vectors for Bluetongue Virus (BTV).

63. De Liberato C, Farina F, Magliano A, Rombolà P, Scholl F, Spallucci V, et al. Biotic and Abiotic Factors Influencing Distribution and Abundance of <I>Culicoides obsoletus</I> Group (Diptera: Ceratopogonidae) in Central Italy. J. Med. Entomol. [Internet]. 2010;47:313–8. Available from: http://openurl.ingenta.com/content/xref?genre=article&issn=0022-2585&volume=47&issue=3&spage=313

64. Conte A, Goffredo M, Ippoliti C, Meiswinkel R. Influence of biotic and abiotic factors on the distribution and abundance of Culicoides imicola and the Obsoletus Complex in Italy. 2007;150:333–44.

65. Lühken R, Gethmann JM, Kranz P, Steffenhagen P, Staubach C, Conraths FJ, et al. Comparison of single- and multi-scale models for the prediction of the Culicoides biting midge distribution in Germany. Geospat. Health. 2016;11:119–29.

66. Kirkeby C, Bødker R, Stockmarr A, Enøe C. Association between land cover and Culicoides (Diptera: Ceratopogonidae) breeding sites on four Danish cattle farms. Entomol. Fenn. [Internet]. 2009;20:228–32. Available from: http://docsrestringidos.cita-aragon.es/monografias/articulos2014/r153_14.pdf

67. Brugger K, Rubel F. Characterizing the species composition of European Culicoides vectors by means of the Köppen-Geiger climate classification. Parasit. Vectors [Internet]. 2013;6:333. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4176262&tool=pmcentrez&rendertype=abstract

68. Kiel E, Liebisch G, Focke R, Liebisch A. Monitoring of Culicoides at 20 locations in northwest Germany. Parasitol. Res. [Internet]. 2009;105:351–7. Available from: http://link.springer.com/10.1007/s00436-009-1409-x

69. Wittmann EJ, Mellor PS, Baylis M. Using climate data to map the potential distribution of Culicoides imicola (Diptera: Ceratopogonidae) in Europe. Rev. Sci. Tech. [Internet]. 2001;20:731–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11732415

70. Mellor PS, Carpenter S, Harrup LE, Baylis M, Wilson A, Mertens PPC. Bluetongue in Europe and the Meditarrenean Basin. In: Mellor P, Baylis M, Mertens PPC, editors. Bluetongue. Academic Press; 2008.

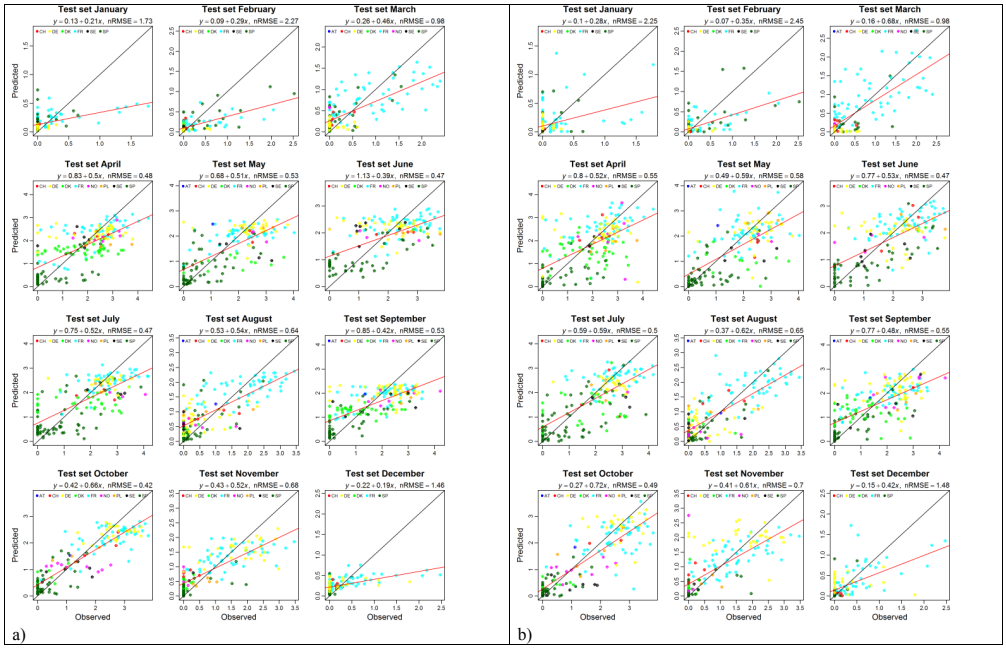
71. Brugger K, Rubel F. Bluetongue Disease Risk Assessment Based on Observed and Projected Culicoides obsoletus spp. Vector Densities. Gubbins S, editor. PLoS One [Internet]. 2013;8:e60330. Available from: http://dx.plos.org/10.1371/journal.pone.0060330

72. Guis H, Caminade C, Calvete C, Morse AP, Tran A, Baylis M. Modelling the effects of past and future climate on the risk of bluetongue emergence in Europe. J. R. Soc. Interface [Internet]. 2012;9:339–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21697167

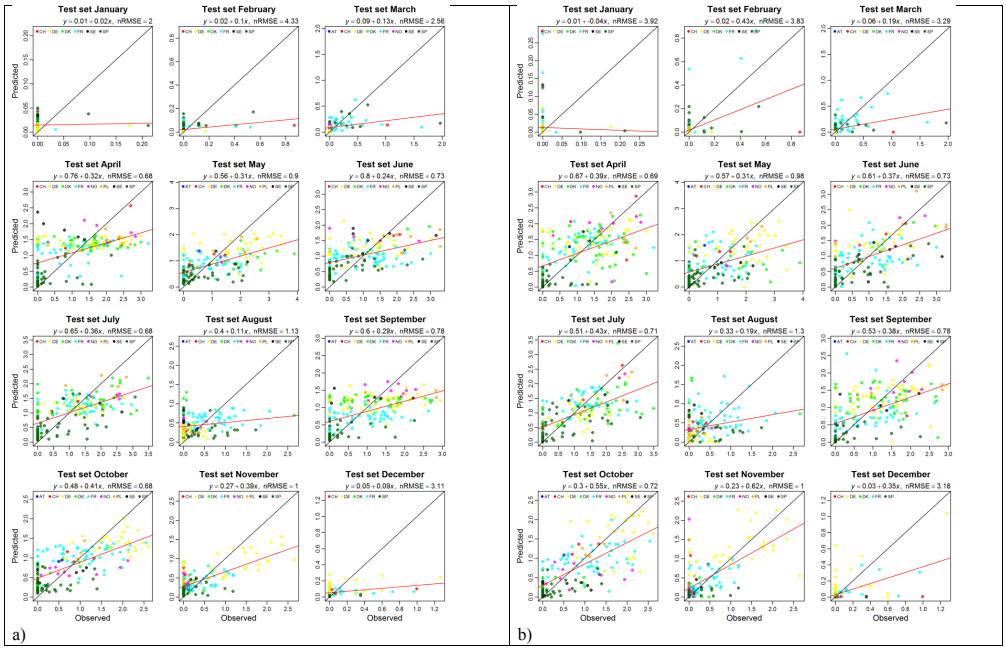
73. Lord CC, Woolhouse ME, Heesterbeek JA, Mellor PS. Vector-borne diseases and the basic reproduction number: a case study of African horse sickness. Med. Vet. Entomol. 1996;10:19–28.

Additional file 1: Figures S1-S7

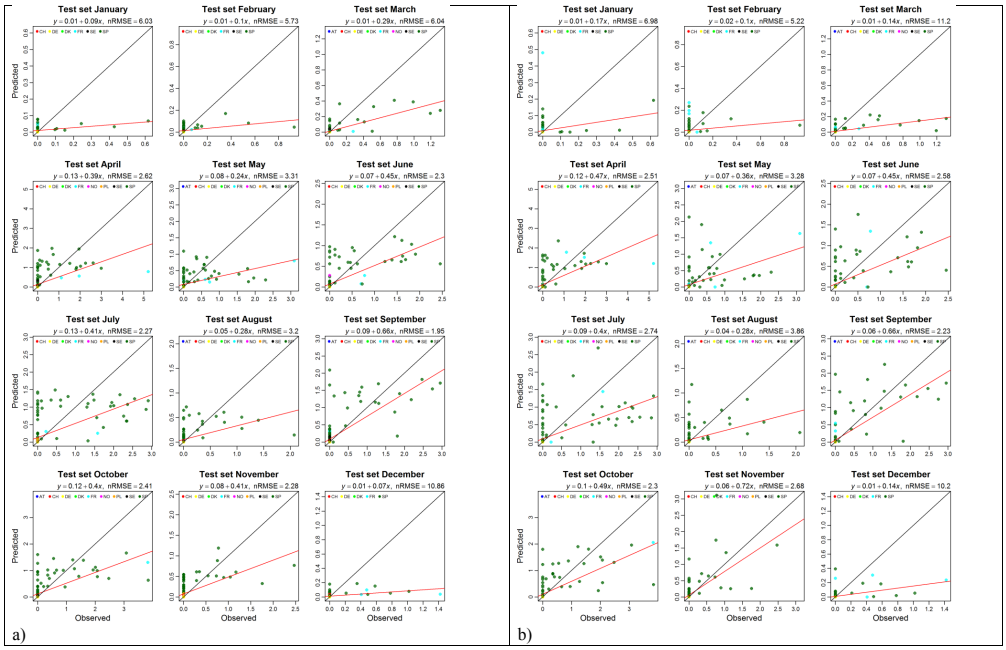
Additional file 1: Figure S1. Plotted observed vs predicted values for the Obsoletus ensemble: a) RF results. b) Interpolation results. Red line: best linear model fit for the predictions. Black line: perfect model fit.



Additional file 1: Figure S2. Plotted observed vs predicted values for the Pulicaris ensemble: a) RF results. b) Interpolation results. Red line: best linear model fit for the predictions. Black line: perfect model fit.

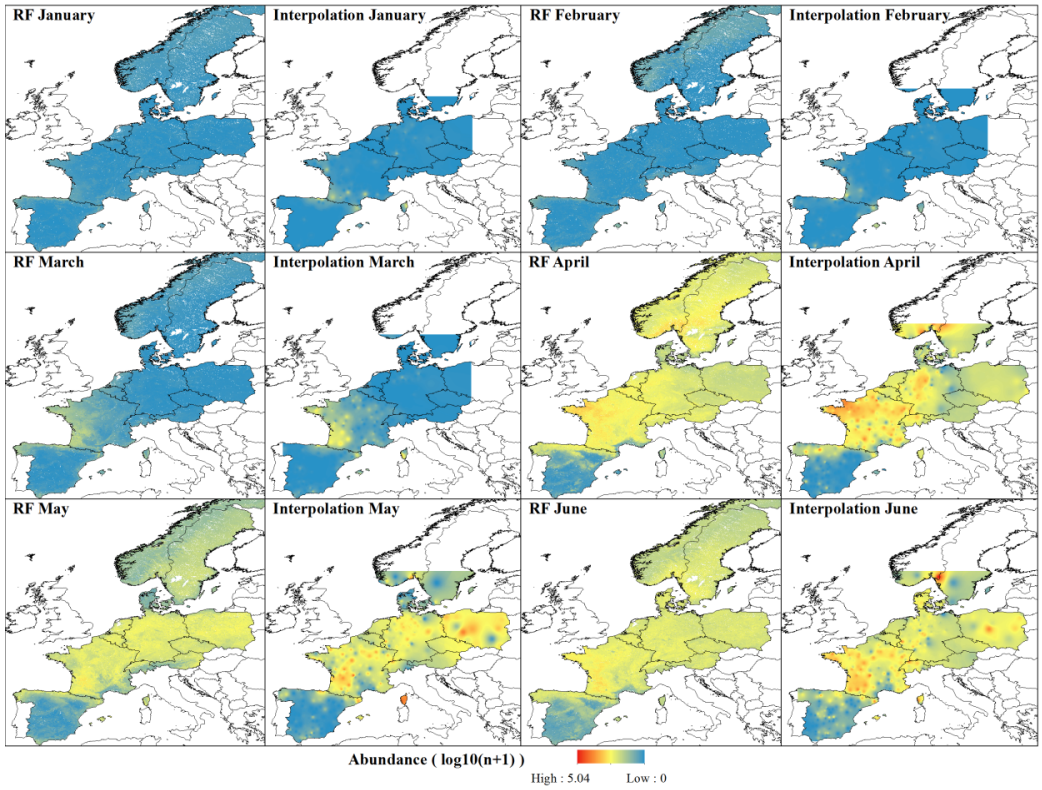


Additional file 1: Figure S3. Plotted observed vs predicted values for *Culicoides imicola*: a) RF results. b) Interpolation results. Red line: best linear model fit for the predictions. Black line: perfect model fit.

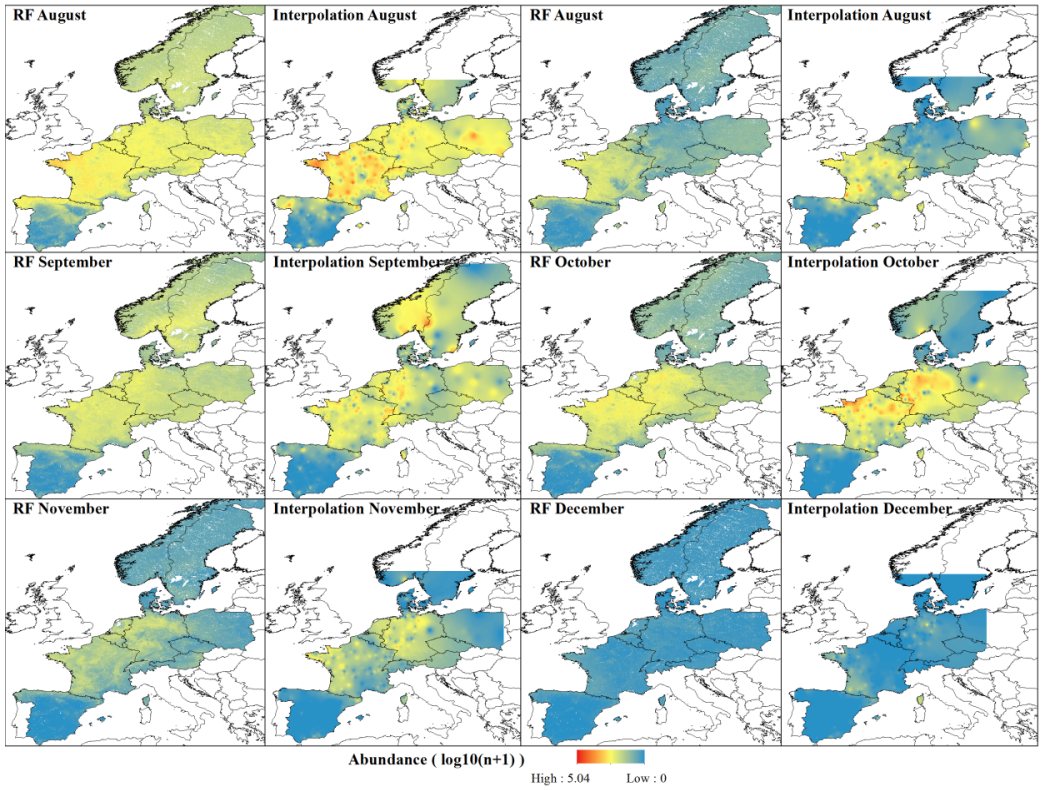


Additional file 1: Figure S4. Comparison of the abundance maps for each month using Random Forest (RF) and Interpolations for the Obsoletus ensemble. a) maps from January to June. b) maps from July to December.

a)

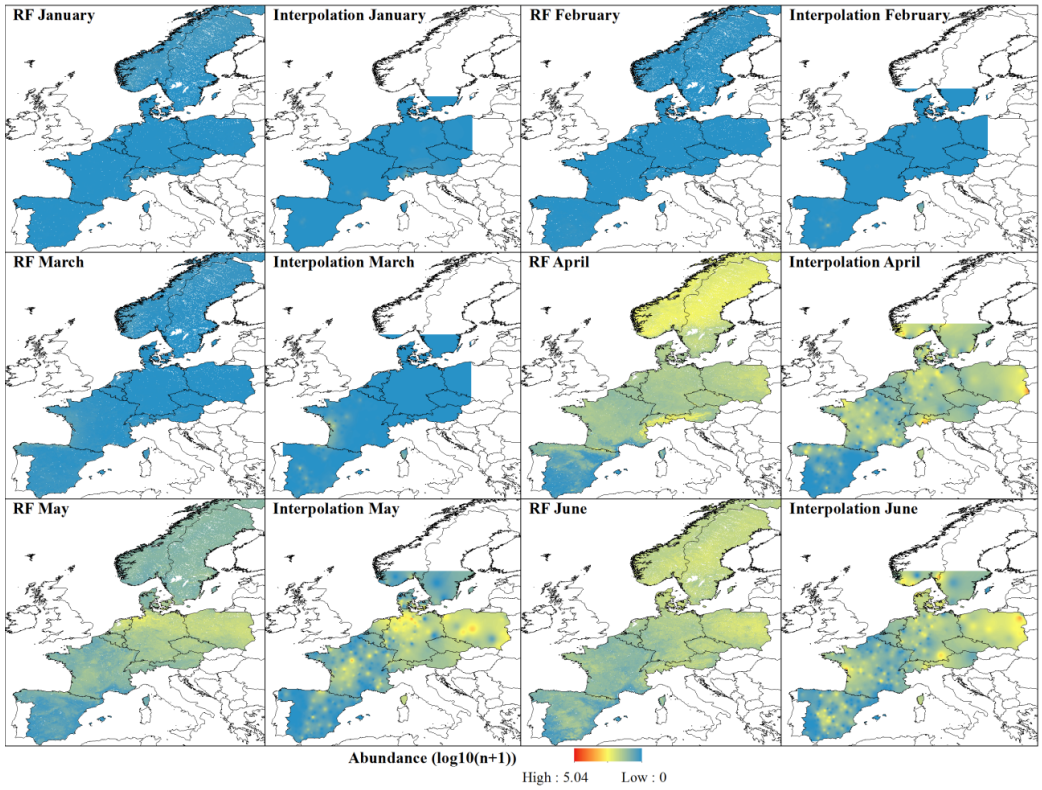


b)

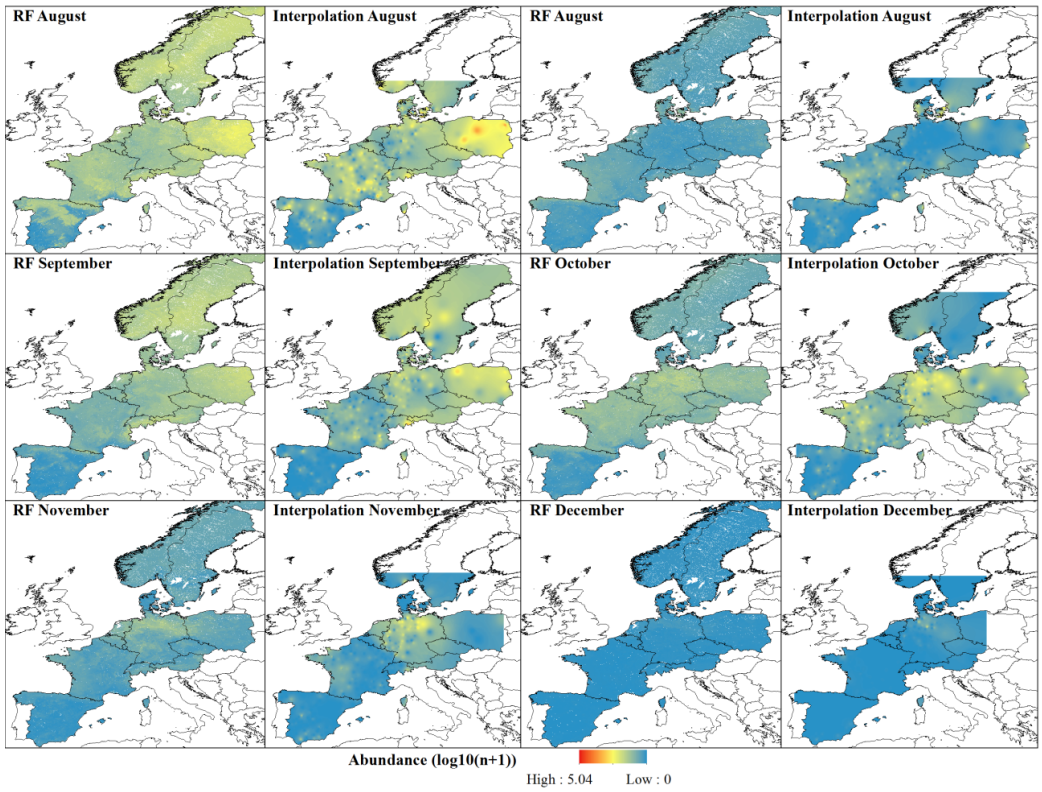

RF August
Interpolation August
RF August
Interpolation August
RF September
Interpolation September
RF October
Interpolation October
RF November
Interpolation November
RF December
Interpolation December
Abundance (log10(n+1))
High : 5.04
Low : 0

Additional file 1: Figure S5. Comparison of the abundance maps for each month using Random Forest (RF) and Interpolations for the Pulicaris ensemble. a) maps from January to June. b) maps from July to December.

a)

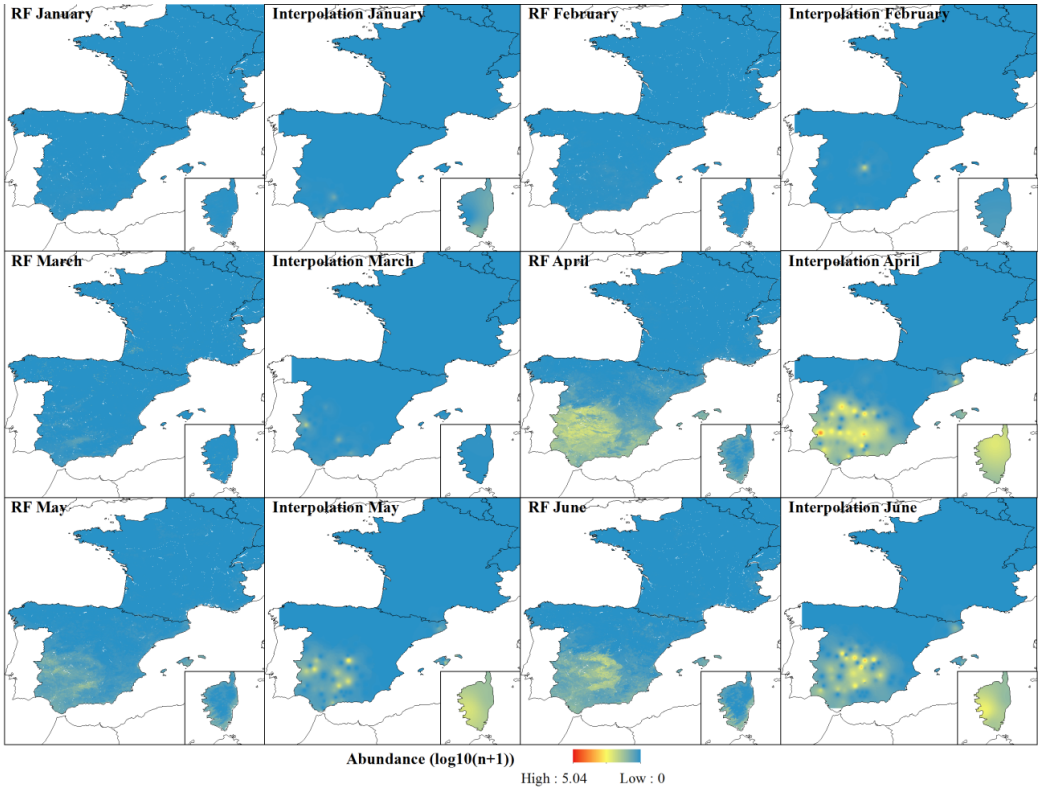


b)

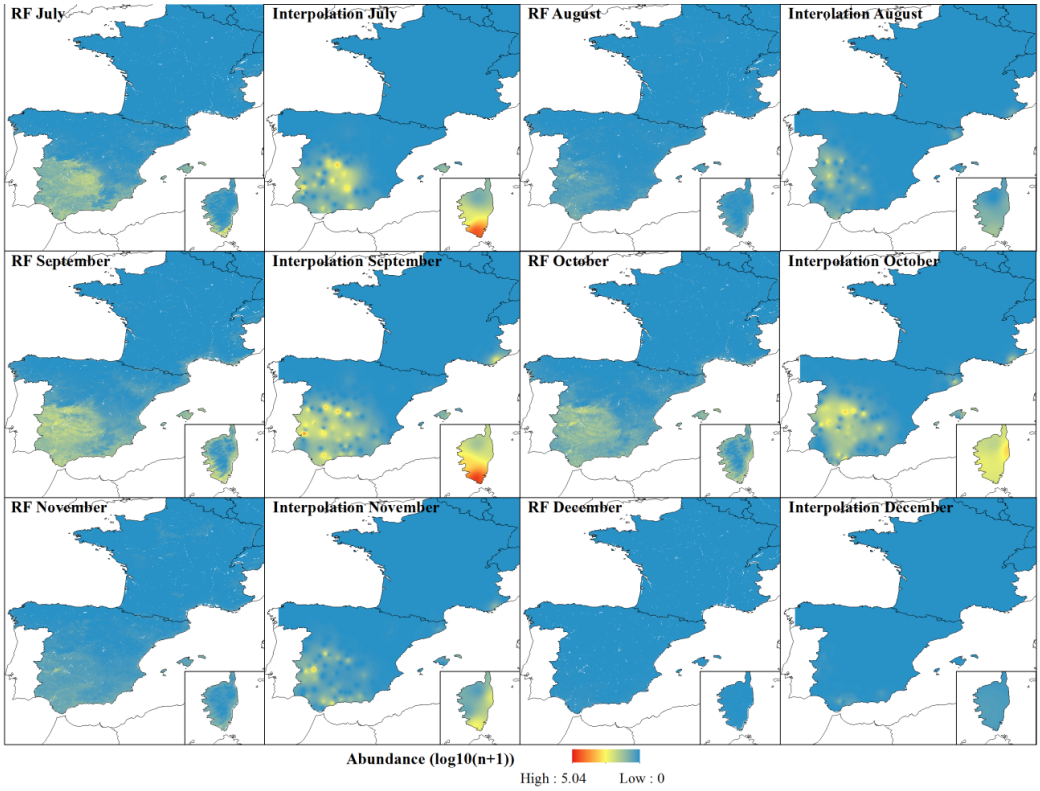

RF August
Interpolation August
RF August
Interpolation August
RF September
Interpolation September
RF October
Interpolation October
RF November
Interpolation November
RF December
Interpolation December
Abundance (log10(n+1))
High : 5.04
Low : 0

Additional file 1: Figure S6. Comparison of the abundance maps for each month using Random Forest (RF) and Interpolations for *Culicoides imicola*. a) maps from January to June. b) maps from July to December.

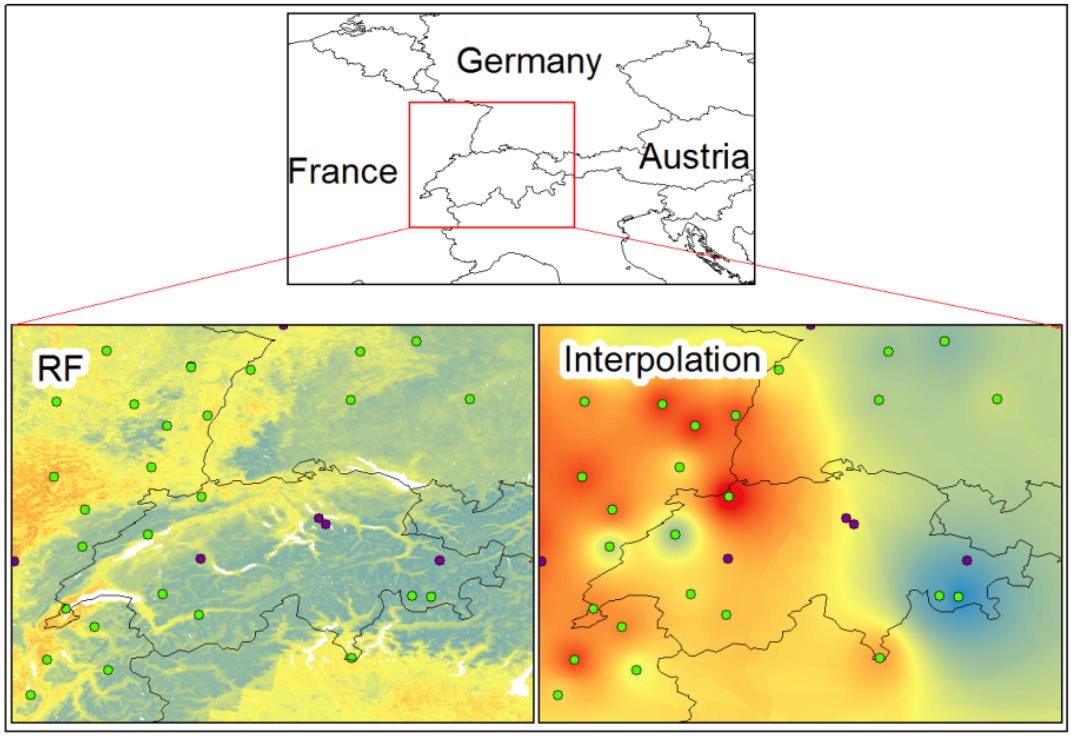
a)



b)


RF July
Interpolation July
RF August
Interolation August
RF September
Interpolation September
RF October
Interpolation October
RF November
Interpolation November
RF December
Interpolation December
Abundance (log10(n+1))
High : 5.04
Low : 0

Additional file 1: Figure S7. At a local scale, interpolation maps produce a smother surface between the farms compared to environmental driven RF, for which the predictions differ between adjacent pixels. The example shown in the figure corresponds to the August maps for the Obsoletus ensemble. Green dots: farms used for training, purple dots: farms within the test set.



3.3.1 Unpublished results relating Manuscript III

Background

Random Forest is a machine learning method which does not require any kind of assumption regarding the nature of the data. Contrary to classical statistic methods such as GLM, modelling using the RF technique do not require an *a priori* knowledge the distribution of the data, neither is it necessary to know if the explanatory variables are correlated or to assume that observations in the training process are independent [Breiman, 2001, Evans et al., 2011]. Because of these characteristics, RF has become a widely used technique in species distribution modelling (SDM) [Cutler et al., 2007], where the available data, often, do not fulfil these requirements. Species distribution modelling analyse the relation between a response variable (presence/absence or abundance) and the explanatory variables measured at given locations and then fit a model, which is later used to predict new observations at un-sampled locations [Elith and Leathwick, 2009]. In this thesis, RF is used to predict the probability of presence (second manuscript) and the abundance (third manuscript) in nine European countries. Despite that RF is used to make spatial predictions resulting in maps, the method is essentially non-spatial and ignores the spatial location of the observations during the training process. The predictions made at un-sampled sites, after modelling, are based only on the predictor's values at those sites, regardless their spatial arrangement and relation between neighbours. Moreover, when estimating model performance through the calculation of the residuals obtained using external validation, the model error is not assessed spatially. Thus, it becomes difficult to determine if there are areas which are better predicted than others [Zhang et al., 2005]. Ideally, in any spatial model the residuals should be randomly distributed in the study area, as spatial correlation among them violates the 'independence of errors' assumption required to justify the use of statistical models in modelling species' distributions.

In this thesis, spatial autocorrelation between the farms used as for training was not analysed as I considered RF to be suitable model for SDM and being capable of producing accurate predictions, even though

some observations close to each other may be correlated. But in this section the spatial correlation of the residuals are analysed. In this section, I present the monthly maps of the residuals obtained for the Obsoletus ensemble, Pulicaris ensemble and *C. imicola*. I also present maps showing the significant spatially autocorrelated clusters (areas with similar values) within the residuals. In an ideal model, the predicted residuals would be expected to be randomly distributed and therefore without significant spatial clusters.

Methods

Calculating the residuals: I used the RF models created for comparing the performance of RF vs the interpolation used in Manuscript III. These models, as explained in the third manuscript, were trained using the monthly average abundance for each farm. Therefore, each farm contained only one record and no predictions are made for individual years. Only one map showing the average abundance was generated per month.

For each month and species group, I used the predicted abundance map from the training datasets to extract the predictions corresponding to each farm of the test datasets and I then calculated the residuals as the observed minus the predicted average abundance. The residuals were plotted on a map. Positive residual values indicate that the model underestimated the average abundance while negative values indicate that the model overestimeted the monthly average abundance.

Cluster analysis of the residuals: To determine if the residuals were spatially autocorrelated, I calculated the coefficient local Moran's I [Anselin, 1995].This step was done using GIS (ArcGIS 10 manufacturer + city and all the usual requirements for software). Local Moran's I compares the values of a pair of neighbour features to the mean value for all the features in the study area and thus detects the presence of spatial dependency (clusters) in some areas. Local methods also returns a p-value for each feature, indicating if clusters are significant different from a random distribution, i.e., identifies areas where the features are significantly correlated ($p < 0.05$) with the features of its neighbours. I used the ArcGIS function "Cluster and Outlier Analysis (Anselin Local Moran's I)" wich produces four new attributes for each farm of the test

set. These are: (i) Local I index, (ii) Z-scores (iii) p-values (iv) cluster type. The attribute “Cluster type” classifies the farms into five classes, each one indicating if the farms are clustered or not, or if they are an outlier. High High category shows points belonging to a cluster of high values, Low Low category shows points belonging to a cluster of low values and categories High Low or Low High show outliers. A fifth "not significant" category indicates that the points are not spatially correlated with its neighbours (i.e there is any spatial correlation) [Mitchell and Minami, 1999].

Results

Obsoletus ensemble: For the Obsoletus ensemble, the residuals appeared to be distributed randomly across the study area with medium residual values (Figure 3.17). The cluster analysis of the residuals showed that for all the months there were no significant cluster, only August, October and November, small clusters of high values appeared in areas of France and Germany 3.18.

Pulicaris ensemble: The monthly maps of residuals for the Pulicaris ensemble showed that the residuals were randomly distributed without any visible cluster (Figure **??**). The cluster analysis of the residuals indicated that, in general there were not significant clusters, except in April, May, August, October and November, where positive cluster were found in Germany, Denmark, France and Spain. Small negative clusters were found in August in Germany and in October in France (Figure 3.20).

Culicoides imicola: For *Culicoides imicola*, the residuals appeared to be equally distributed. However, during April, May and October, there were a few farms with high residual values in Spain. Cluster analysis identified these extreme values as outliers for April and October, but not for May, where they were identified as forming a cluster of high values. Others high values clusters were identified in June and July (Figure 3.21). A cluster of low values was found in April and September in Spain. Apart from these clusters, the residuals were not significantly locally spatially correlated.

For *Culicoides imicola*, the residuals appeared to be equally distributed. However, during April, May and October, there were a few farms with high residual values in Spain. Cluster analysis identify these

extreme values as outliers for April and October, but not for May, where they were identified as a cluster of high values. Others high values clusters were identified in June and July (Figure 3.21). A cluster with low values was found in April and September in Spain. Apart from these clusters, the residuals were not correlation as (not significant clusters).

Discussion

For this thesis, any cluster analysis was performed on the training set farms, despite the fact they are expected to be autocorrelated as Tobler's First Law of Geography states: "Everything is related to everything else. But near things are more related than distant things". Here, I assume that RF would be able to pick differences in the ecological niche from the predictor variables and the models were run without accounting for the spatial relation within the training data. RF is a essentially a non-spatial model approach so therefore, the RF models for predicting the abundance were trained without any spatial input, considering only the relation between the response and the environmental features measured at given points. Even so, it is expected that the residuals would show a random distribution, i.e. no clustering.

I assessed the spatial distribution of the residuals for the Obsoletus and Pulicaris ensemble and for *C. imicola*. In general, the residuals were randomly distributed in space and only small clusters of residuals were found for Pulicaris ensemble and for *C. imicola* and almost none for the Obsoletus ensemble.

Despite of the ability of the method to handle spatial data, some researchers have tried to incorporate the spatial component in the prediction process. For instance, Hengl et al. (2018) presented a RF "for spatial predictions framework (RFsp) where buffer distances from observation points are used as explanatory variables, thus incorporating geographical proximity effects into the prediction process" [Hengl et al., 2018]. Another work, has bind together boosted regression trees together with kriging, and they concluded that using BRT was sufficient to achieve accurate predictions [Martin et al., 2014]. It would be interesting to analyse the present data with models accounting for spatial correlation, to see if there is adding spatial information may improve the present abundance predictions.

Conclusion

Because RF is a method which does not require independent observations, it was a suitable method to model the *Culicoides* abundance used here, as abundance is likely to be spatially autocorrelated. The fact the residuals were not correlated indicates that RF was able to deal with spatial autocorrelated data and the variation found in the residuals was not subjected to certain region of the study area, but product of the factors inherent to the model.

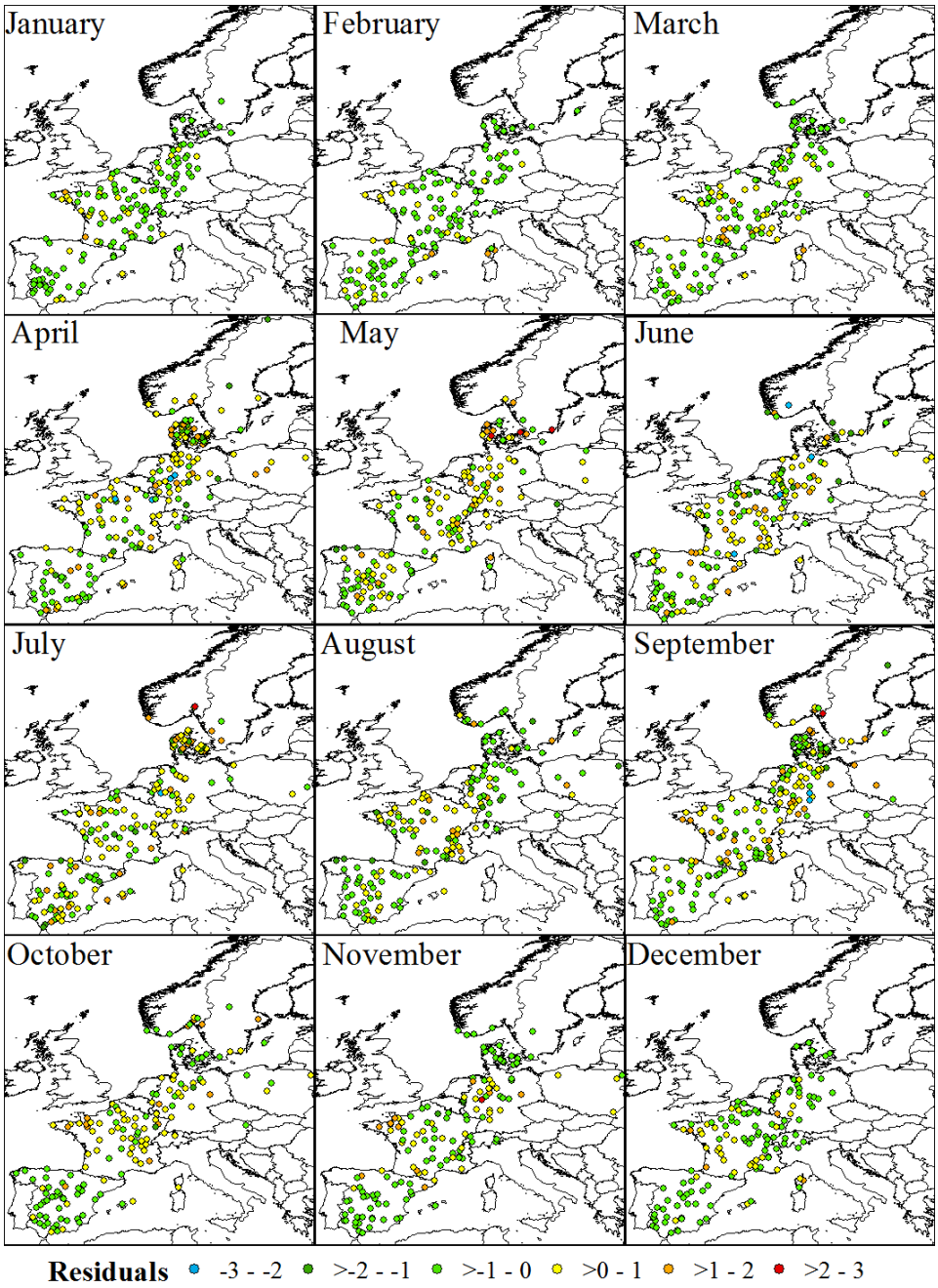


Figure 3.17: Residuals (expressed in log10(n+1)) between the observed and the predicted average abundance for the Obsoletus ensemble. They are expressed in log10(n+1)

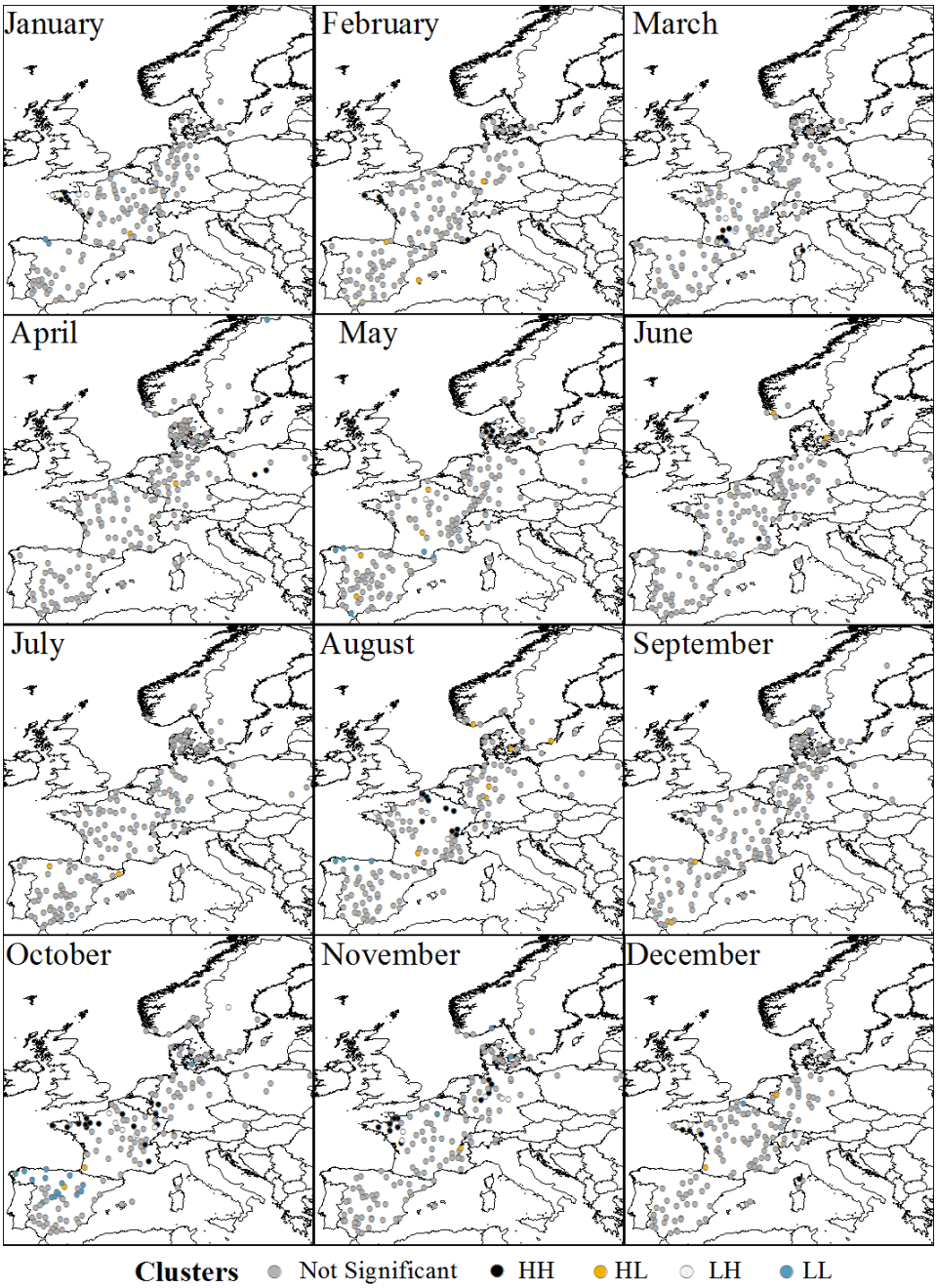


Figure 3.18: Cluster analysis (based on local Moran's index) of the residuals between the observed and the predicted average abundance for the Obsoletus ensemble. HH: High High, HL: High Low; LH: Low High; LL: Low Low.

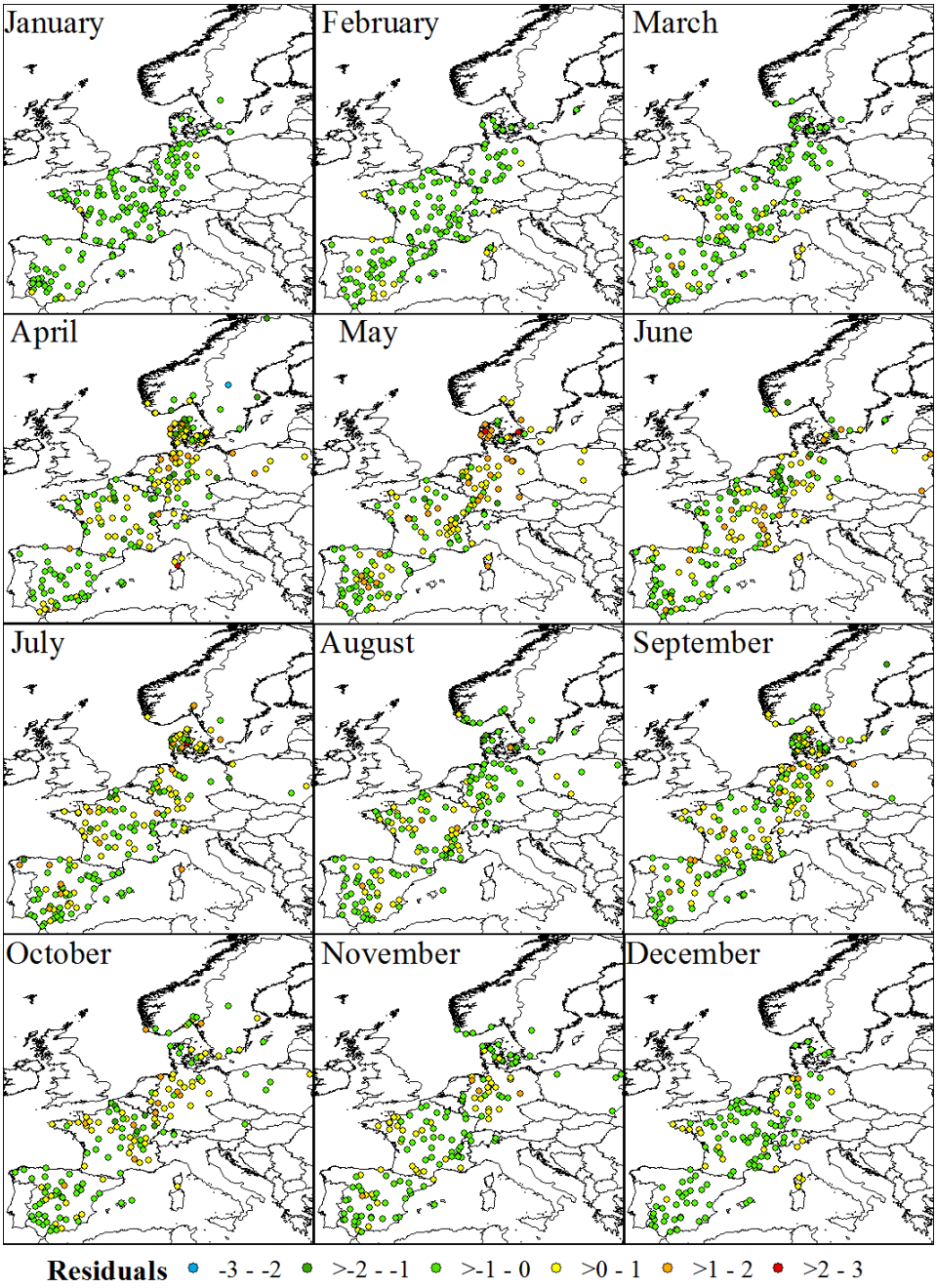


Figure 3.19: Residuals (expressed in $\log_{10}(n + 1)$)between the observed and the predicted average abundance for the Obsoletus ensemble.

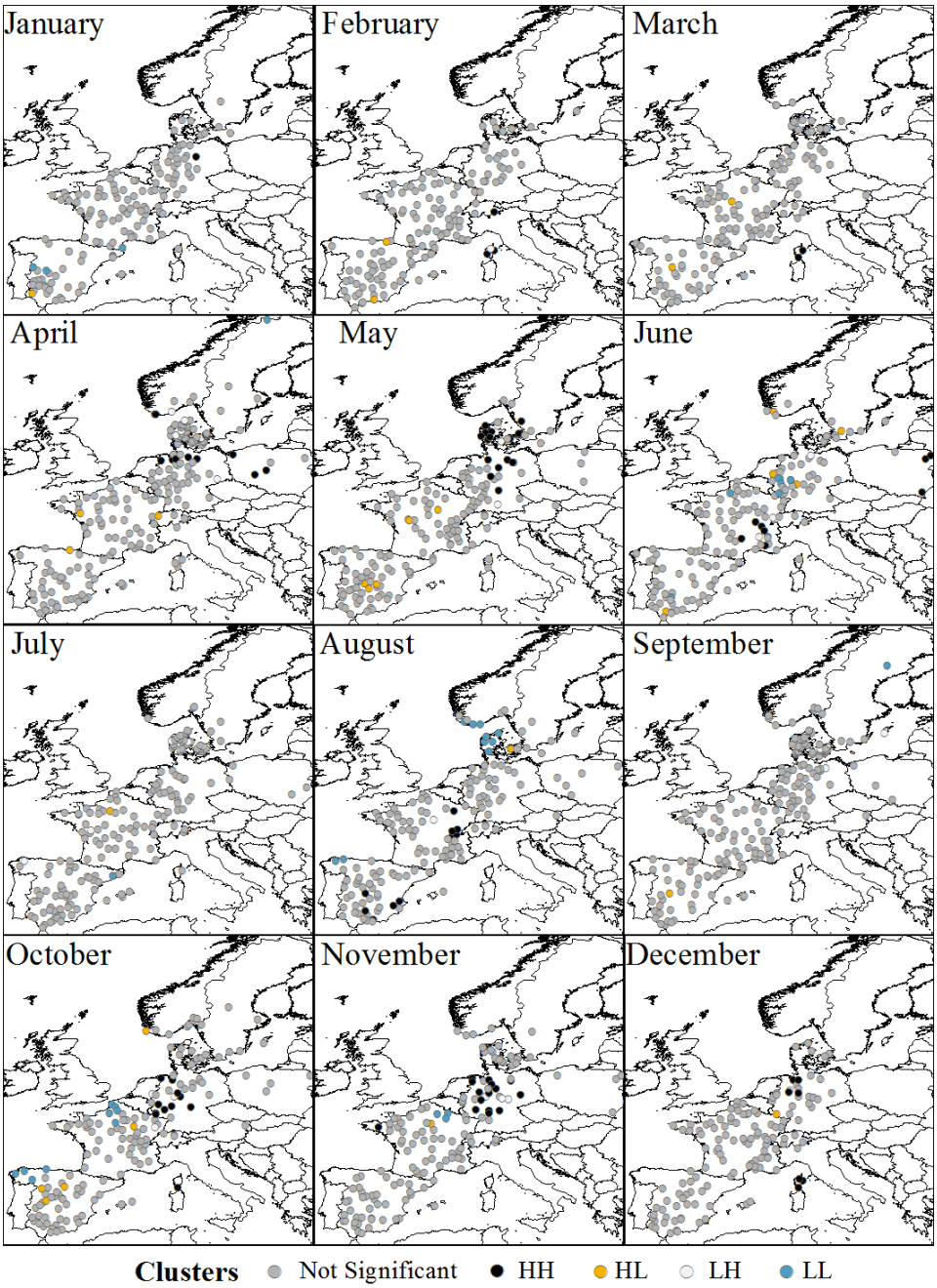


Figure 3.20: Cluster analysis (based on local Moran's index) of the residuals between the observed and the predicted average abundance for the Pulicaris ensemble. HH: High High, HL: High Low; LH: Low High; LL: Low Low.

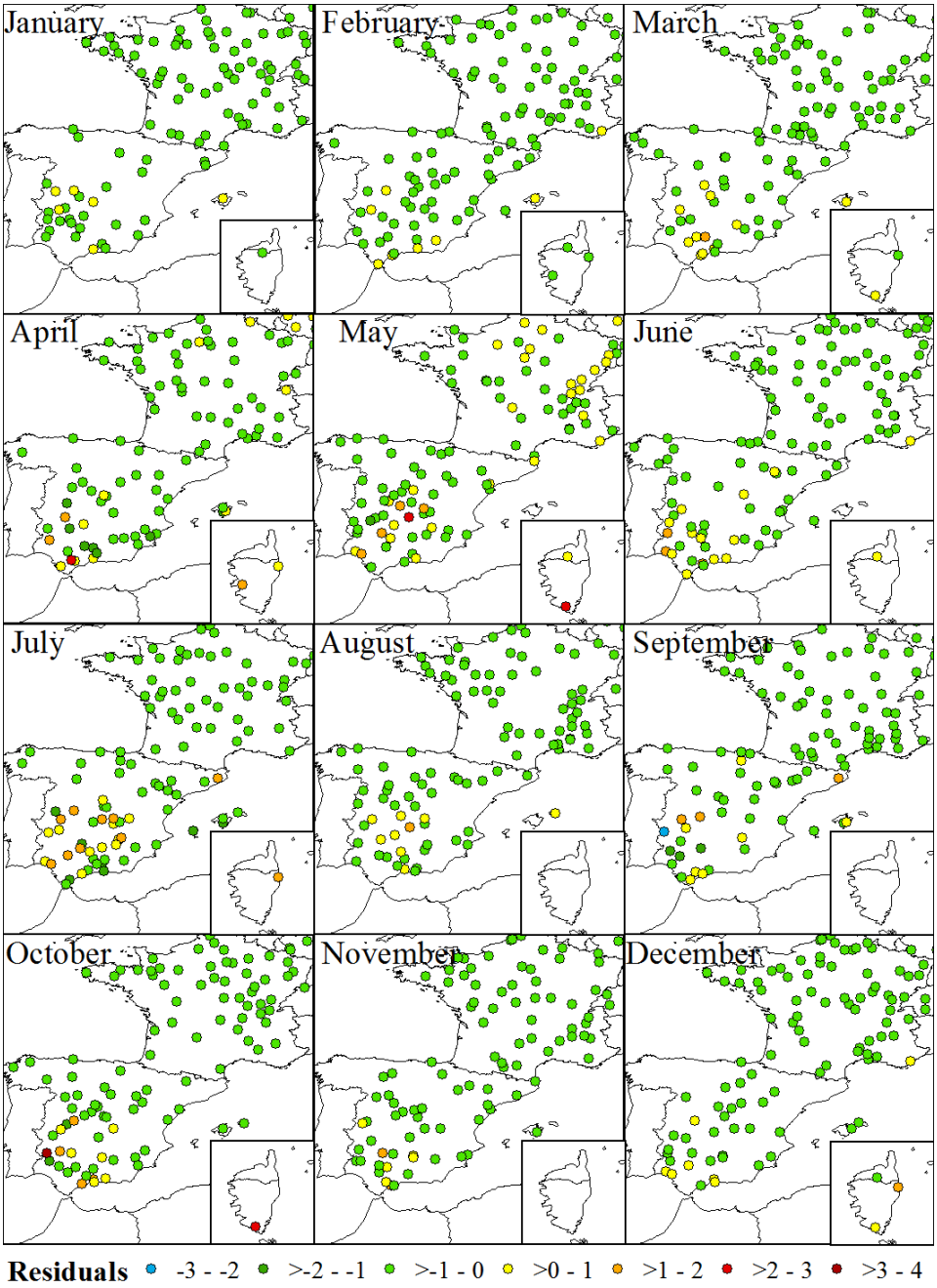


Figure 3.21: Residuals (expressed in $\log_{10}(n + 1)$)between the observed and the predicted average abundance for the Pulicaris ensemble.

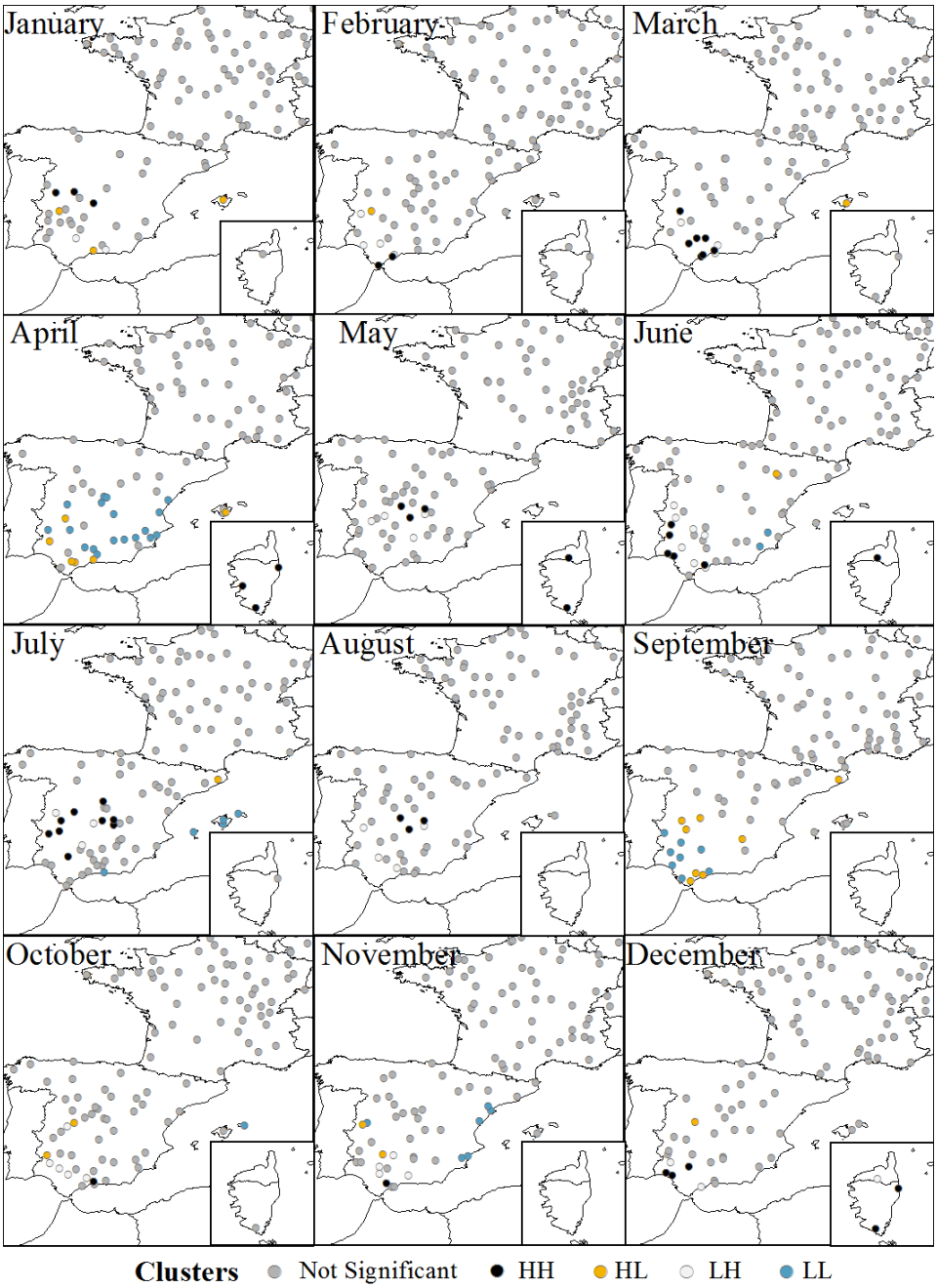


Figure 3.22: Cluster analysis (based on local Moran's index) of the residuals between the observed and the predicted average abundance for *C. imicola*. HH: High High, HL: High Low; LH: Low High; LL: Low Low.

Chapter 4

Discussion

The overall aim of this thesis was to understand the spatio- temporal distribution of the main *Culicoides* vectors of bluetongue and Schmallenberg in Europe, at a nearly continental scale. To do so, we used entomological data collected in nine European countries and predicted the occurrence and abundance for *C. imicola*, and for Obsoletus and Pulicaris ensembles for a large area of Europe, from southern Spain to Northern Sweden [Cuéllar et al., 2018].

A descriptive analysis of the observed data was performed (Manuscript I, see section 3.1). We compared the abundance data aggregated by every latitude range of 5° latitude, mapped the observed abundance per month (to detect spatio-temporal patterns) and analysed the start of the vector season at NUTS 3 level. One of the main findings, was that the start of the vector activity began later at higher latitudes. This is expected as optimal temperatures would arise later in the year at higher latitudes but we observed that the start of vector season occurred, at high latitudes, with mean temperatures for the region as low as 1° C (much lower temperatures than at the south) for the Obsoletus ensemble. This results showed that the vector activity can occur with colder temperatures than expected (which is 8-10° C [Purse et al., 2007, Versteirt et al., 2017]) and that the start of the vector season cannot simply be predicted from the daily temperatures at spring. Hence, temperatures cannot be used as a simple inexpensive proxy for the start of the vector season in Europe. This is an interesting finding because the first definitions of the so-called 'vector free season' developed by the EU Commissions were based on temperatures in areas where vector

surveillance were not operating (10° C) [EFSA, 2017].

Another interesting result of the descriptive analysis was the extreme high abundance (more than 80.000 per night observed at some farms at high latitudes, such as the southern coast of Norway. High Obsoletus ensemble abundances (more than 80.000 per night) were reported in France [Balenghien et al., 2010] and Germany [Mehlhorn et al., 2009] but this was the first the time that an extreme high abundance is reported from southern coast of Norway. It is important to notice that these extreme high abundances found in farms in the south of Norway were not common and therefore, it would be interesting to investigate possible causes for these outliers (or whether they are outliers) and examine available environmental data previous to their occurrence. It might be possible that an extreme event (extremely high temperatures, for instance) or unusual farm practices could have affected the dynamics of Obsoletus ensemble specimens for those particular farms.

We predicted *Culicoides* occurrence and abundance to un-sampled areas in Europe using the machine learning technique Random Forest (RF) [Breiman, 2001], a technique that has been proven to outperform other classic statistical modelling techniques in test comparisons [Cianci et al., 2015, Cutler et al., 2007, Van Doninck et al., 2014]. RF can be used to predict a class or probability of class occurrence, when modelling the response variable as occurrence data (Presence or Absence data), or RF can be used to predict a numerical value, if the input data is numerical (e.g. abundance) [Breiman, 2001, Liaw et al., 2002]. In Manuscript II (see section 3.2, we showed the results of using RF for predicting vector occurrence data and mapped the predictions for each month of the year. Manuscript III (section 3.3) shows the results of using RF on the same data set, but using abundance as input data. We presented abundance maps for each month of the year.

Modelling *Culicoides* data

The performance of RF varied according to the nature of the input data (i.e. if it was used as occurrence or abundance data). In the case of occurrence data (manuscript II) RF forest performed very well for *C. imicola*, well for Obsoletus ensemble and fair for Pulicaris ensemble.

Considering abundance at farm scale, the results showed that RF had

a fair performance for Obsoletus ensemble, for Pulicaris ensemble and *C. imicola*, even though the geographic pattern found in the resulting maps matched with the known of the distribution of Pulicaris and *C. imicola* (Manuscript III, section 3.3).

It is difficult to predict abundance, even in areas having the same environmental conditions [VanDerWal et al., 2009]. For example, within the dataset, we observed farms placed in the same 1 km pixel having a large difference in abundance among them (data not shown). The results in Manuscript III (section 3.3) highlight the difficulties found for predicting vector abundance. The only fair performance of the RF might be due to:

1. high inter-annual variation found in the data: the monthly mean abundance varied among years. Inter-annual variation in abundance is not uncommon and has been noticed in previous studies [Brugger et al., 2016, De Liberato et al., 2010]. Inter annual differences in the vector abundance may be due to weather conditions (for instance an unusual hot or rainy season) or might be the result of different generations through the vector season. Different generations have been reported in a year for Obsoletus ensemble [Balenghien et al., 2010, Foxi et al., 2011, Hill, 1947] and *C. imicola* [Braverman and Linley, 1988] and the timing of the generations peak does not occur at same time of the year in different years, as shown by the Danish *Culicoides* surveillance system (`www.myggetal.dk`).

2. the environmental predictors used here are not the ones driving the vector abundance: even though the predictors used here have been reported to have an effect on *Culicoides* abundance [Calvete et al., 2006, Conte et al., 2007, Purse et al., 2015, 2007, Tatem et al., 2003] other predictors not considered here might have been the main drivers of *Culicoides* abundance. For example, farm practices are a set of variables that may affect *Culicoides* abundance by favouring the presence of suitable breeding sites. *Culicoides* spp breed on different type of substrates, with organic matter and enough moisture for the larvae to develop [Kettle, 1962]. In northern Spain, *C. obsoletus* was found breading in different types of manure [González et al., 2013], similar to Germany where it was

associated to dung heaps [Steinke et al., 2016]. Farm management practices are a variable that cannot be measured by remote sensing and therefore, could not be considered for this work.

3. the spatial resolution of 1 km was not optimal to capture certain landscape features that might be indicators of suitable breeding sites in the surroundings of the farms. Drainage canals, muddy substrates, surfaces neighbouring leaking pipes have been reported to be suitable as breeding sites for *C. imicola* [Braverman and Linley, 1988, Foxi et al., 2011, Mellor and Prrzous, 1979].

To my knowledge, there are only a few studies from Europe that have modelled the *Culicoides* abundance and mapped the predictions (without extrapolating to other regions). These studies are for *C. imicola* in Spain [Acevedo et al., 2010, Ducheyne et al., 2013] and *C. impunctatus* in Scotland [Purse et al., 2012]. European Union's decisions on control and management of possible outbreaks are based on joint decisions among the member states and therefore, there is a need to produce abundance maps for Europe [EFSA, 2017], that can be used to improve risk assessments analyses based on R_0 modelling [Hartemink et al., 2015] at a continental scale and to distribute EU funding for surveillance and control in a fair and transparent manor. The abundance maps presented here constitute the first abundance maps made for Europe till date and, even though vector abundance at individual farm level could not be accurately predicted, the maps are still useful at identifying regions having different environmental features and different abundances and therefore different risk of transmission within Europe.

Random Forest vs interpolation

In Manuscript III (section 3.3, the performance of RF models for predicting vector abundance was compared the performance of a simple interpolation method (IDW algorithm). For interpolation is not necessary to have predictor variables as interpolation only considers the position of the sampled locations and the abundance itself for making a prediction to non-sampled areas. Interestingly interpolation performed only slightly poorer compared to RF. This result highlights how a sophisticated predictor driven technique like RF was not able to

completely outperform a simple interpolation method [Braverman and Linley, 1988].

This raise the question: could interpolation be used instead of RF?

Even though interpolation is a straightforward method that allows mapping without using any predictor variable, IDW method still has disadvantages such as the lack of rules of thumbs to calculate the value of the optimal parameters (i.e. power, minimum and maximum neighbour's numbers). More importantly interpolation depends on the amount of data points available as predictions are made using the neighbouring points. It is expected that the errors obtained from interpolation would vary spatially: the higher the density of points within a region, the more accurate a prediction would be within that region. Spatial autocorrelation is necessary for the method to predict correctly.

RF is essentially a non-spatial method and uses a different predictive approach as the model outputs are based on suitable ecological niches (we need environmental variables as predictors) and therefore, RF predictions are not depending on the point density in a given area. I do not know if RF in practise is capable to perform worse than an interpolation method and, as vector abundance is known to have complex relation with environmental variables, RF should be preferred rather than interpolation. RF should be a suitable method to find complex relations between the response variable and predictors in most of the cases.

I would like to remind the reader about some of the advantages of using Random Forest: it has been proven to outperform classical statistical methods (which we could not have used in this thesis because the data does not full fill the normality criteria and there is very strong colinearity between predictors); within the machine learning techniques, is simple to tune, requires less computing time and the process can be run in parallel which decreases, in some cases, the computing time. In the data set gathered for this project, there were 904 sampled points (farms) distributed more or less uniformly across the nine studied countries (an exception was Denmark, where more than 300 farms were sampled and Sweden, in which the sampled farms were clustered south and a few farms were sampled northern wards).

Challenges

Modelling vector occurrence/abundance data originating from different resources is challenging. There is no rule of thumb on how to model abundance data: some modellers take the maximum catch as the best measurement of abundance [Calvete et al., 2006, Ducheyne et al., 2013, Purse et al., 2004]. In this thesis we used the average abundance because we acknowledge that there might be extreme observation that do not represent the real abundance in the area (for example the farms in southern coast of Norway in our dataset, (data not shown)). The final decision about to model should be according to the objective proposed.

Regarding the predictors, it is possible that there are other factors driving *Culicoides* abundance at a farm scale. For instance, farm management practices creates the suitable conditions for breeding environments [Carpenter et al., 2009, Venail et al., 2012]. Such information should be obtained when trapping is carried out in each farm and if used as a predictor variable, it might be able to improve the model performance.

Obtaining satellite images with higher resolution might be another possible option to improve the predictions. The use of high resolution images would make it possible to identify landscape features indicating the presence of suitable breeding sites at very local scale (for example SPOT satellite imagery are able at spatial resolution of 10 m). However, using high resolution imagery might be problematic to carry out large scale analysis, as it becomes very difficult (if not impossible) to create a single predictor image covering the extent of the nine countries studied here. Nevertheless, this idea might work for making predictions at smaller scale, such as districts.

Other modelling techniques can be applied to this dataset. For this thesis, I also tried other machine learning techniques such as Boosted regression trees [Elith et al., 2008] and M5 trees (explained in Kuhn and Johnson [2013]) (data not shown) in order to explore if these methods would be able to improve the abundance predictions. The results were not better than the ones based on RF presented in the Manuscripts II and III.

Model performance for abundance was assessed using external validation (i.e. RF model is used to predict the abundance in samples of an independent test set). The dataset used here comprises Culicoides collections only taken in cattle, sheep and horse farms and no other land

cover type was sampled. Therefore, when we calculated the nRMSE to determine the model performance, we are assessing how good the model was to predict abundance for a single land cover, in this case farms. The performance of the model for predicting abundance in other land cover types is not known, as there is no data to validate the predictions. Thus, it is a possible that the RF maps might have been much better for predicting vector abundance at different land covers other than the sampled at farms. This has been the case for mosquitoes, that are a risk to humans everywhere, not only at farming areas RF succeeded to predict mosquitoes abundance in the Netherlands throughout different habitats, based on similar predictors as ours [Ibañez-Justicia and Cianci, 2015].

The approach used in this thesis to model the probability of presence and abundance of *Culicoides* was purely statistic. In machine learning the relations between the response and explanatory variables is analysed but not for inference purposes but to fit a model able to make a predictions on a new dataset. This approach does not require any biologic information of any kind regarding the mechanisms that underlies among the response and predictors. On the other hand, biological models take into consideration biological aspects, for example life stages, fecundity and/or demographics and seasonal dynamics and try to infer the relation found between the response and explanatory variables. For this thesis, to fit a model including fecundity of the *Culicoides* females is not possible, simply because we did not have that information available from all countries. On the other hand, models incorporating population dynamics might be more suitable for modelling the vector abundance [Rigot et al., 2012]. Here, we modelled the *Culicoides* abundance through an average year, considering each month as independent dataset. A model including population dynamics might help to improve the predictions of the mean abundance for Palearctic Culicoides and for *C. imicola* in Europe as it takes into consideration temporal correlation that might exist between different months.

The predictions obtained from the RF models are static, in the sense that they cannot be used to predict a possible response under an environmental change scenario. The reason is that the relation between explanatory variables and the response is not known and probably very

complex, and therefore if the predictor variables change (as it would happen if the environment changes) it is not possible to know *a priori* how the change would affect the response. For instance, changing the mean temperature will have an impact in the *Culicoides* abundance, but the magnitude and the direction of the change will be unknown.

Chapter 5

Conclusions & Perspectives

Conclusions

The main conclusions of this thesis can be summarized as follows:

1. **Descriptive analysis of observed collected data:**

   Obsoletus ensemble:

   - The Obsoletus ensemble was widely distributed and found in high abundance in farms of France, Germany and occasionally southern Scandinavia.
   - There was a spatial latitudinal trend in which the highest accumulated number of biting midges increased towards the north despite the length of the vector period decreased towards the north.
   - The vector season started later at northern latitudes (as expected) but interestingly the vector activity started at a mean regional temperature of 1 C much lower than the starting temperature at southern latitudes. It is therefore not possible to use temperature as a simple proxy for the end of the vector free period in Europe as it has been done in EU regulations during the BTV8 outbreak.
   - Extreme high abundances were reported in southern coast of Norway at two farms (> 80.000 specimens per night). More surveillance in the area is needed to determine is the area can be considered as a high abundance area.

Pulicaris ensemble:

- The Pulicaris ensemble was also widely distributed and the highest abundance farms were located in Germany and Poland. Its abundance was 10-fold lesser than Obsoletus ensemble abundance.
- As the Obsoletus ensemble, there was a spatial latitudinal pattern in which the abundance increased towards but, contrary as Obsoletus ensemble, it decreased again at northern latitudes.

***Culicoides imicola*:**

- *Culicoides imicola* was found only in central and southern Spain, the Var department in France and Corsica. It was much less abundant than the Obsoletus ensemble.
- High abundance farms were found in central and southern Spain and highest peak abundance was recorded in a farm in Corsica.
- For all of the species groups the season started earlier in southern latitudes.

2. **Predicting *Culicoides* occurrence and classifying Absence, Uncertain and Presence classes:**

- It was possible to predict presence and absence of vectors in Europe.
- The Random Forest machine learning technique performed well for the Obsoletus ensemble, fair for the Pulicaris ensemble and very well for *C. imicola*.
- We were able to identify areas and months where vectors were absent or present. For many months these areas constitutes a relatively large proportion of Europe. This allows entomological surveillance to be focused on relatively limited areas thus potentially reducing cost of surveillance significantly. However, limitations of the model performance should be addressed and expert knowledge should be considered to complement decision making regarding animal movement restriction of implementation of surveillance programs.

3. **Predicting *Culicoides* abundance:**

- It was possible to predict monthly abundance of *Culicoides* vectors in Europe.
- When predicting abundance the Random Forest performance varied according the season with a lower predictive power for winter compared to summer months.
- RF was able to make predictions corresponding to different regions in Europe. For instance, the model predicted higher abundance in Germany and France compared to Spain, but it failed to predict *Culicoides* abundance at farm scale.

Perspectives

Abundance is not easy to predict as large spatial variation is found in apparently homogenous environments.

Here, I explored Random Forest machine learning technique. This method has been shown to perform well in several ecological studies. In Manuscripts II and III, I show that RF had a fair performance for predicting the abundance at farm scale and some considerations are necessary for the improvement of the predictions:

- More possible predictor variables should be explore, such as soil conditions which has been proven to affect *Culicoides* abundance.
- For predicting vector abundance in single farms, farm management practices should be measured and added as possible predictors.
- High resolution image may also improve model predictions, as they are able to give information regarding local features at farm scale that might affect vector abundance.
- Other machine learning techniques are available and might be a better option for predicted abundance. Other techniques other than Boosted Regression Trees and M5 trees were not tried.
- Models including population dynamics could be tried on this temporal dataset, as it has been seen that it is possible to achieve good results for predicting abundance. The temporal fluctuations of a population depends (sometimes) on previous population sizes. It could be very interesting to analyse the effect of population size of previous months.

Bibliography

P Acevedo, F Ruiz-Fons, R Estrada, A L Márquez, M A Miranda, C Gortázar, and J Lucientes. A broad assessment of factors determining Culicoides imicola abundance: modelling the present and forecasting its future in climate change scenarios. *PloS One*, 5(12):e14236, 2010.

A Afonso, J C Abrahantes, F Conraths, A Veldhuis, A Elbers, H Roberts, Y Van der Stede, E Méroc, K Gache, and J Richardson. The Schmallenberg virus epidemic in Europe—2011–2013. *Preventive Veterinary Medicine*, 116(4):391–403, 2014.

M Ander, R Meiswinkel, and J Chirico. Seasonal dynamics of biting midges (Diptera: Ceratopogonidae: Culicoides), the potential vectors of bluetongue virus, in Sweden. *Veterinary Parasitology*, 184(1):59–67, 2012.

Luc Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.

T Balenghien, J C Delécolle, and I. Rakotaoarivony. Bluetongue - Report on Entomological Surveillance in France in 2010. *Bulletin Épidémiologique, Santé Animale et Alimentation*, 46:26–31, 2010.

G E Batista, R C Prati, and M C Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.

M Baylis, PS Mellor, EJ Wittmann, and DJ Rogers. Prediction of areas around the mediterranean at risk of bluetongue by modelling the distribution of its vector using satellite imaging. *Veterinary Record*, 150 (13):404–404, 2002.

A Borkent. *The subgeneric classification of species of Culicoides — Thoughts and a warning*. New York: Bulletin of the American Museum of Natural History, 2010, 2014.

Y Braverman and JR Linley. Parity and voltinism of several culicoides spp.(diptera: Ceratopogonidae) in israel, as determined by two trapping methods. *Journal of medical entomology*, 25(2):121–126, 1988.

L Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

K Brugger and F Rubel. Bluetongue disease risk assessment based on observed and projected Culicoides obsoletus spp. vector densities. *PLoS One*, 8(4):e60330, 2013.

K Brugger, J Köfer, and F Rubel. Outdoor and indoor monitoring of livestock-associated culicoides spp. to assess vector-free periods and disease risks. *BMC Veterinary Research*, 12(1):88, 2016.

C Calvete, M A Miranda, R Estrada, D Borras, V Sarto i Monteys, F Collantes, JM Garcia-de Francisco, N Moreno, J Lucientes, EJB Veldhuis Kroeze, et al. 1195801. spatial distribution of culicoides imicola, the main vector of bluetongue virus, in spain. *The Veterinary Record*, 158(4):130–131, 2006.

S Caracappa, A Torina, A Guercio, F Vitale, A Calabro, G Purpari, V Ferrantelli, M Vitale, and PS Mellor. Identification of a novel bluetongue virus vector species of Culicoides in Sicily. *The Veterinary Record*, 153(3):71–74, 2003.

S Carpenter, A Wilson, and P S Mellor. Culicoides and the emergence of bluetongue virus in northern Europe. *Trends in Microbiology*, 17(4): 172–178, 2009.

S Carpenter, M H Groschup, C Garros, M L Felippe-Bauer, and B V Purse. Culicoides biting midges, arboviruses and public health in Europe. *Antiviral Research*, 100(1):102–113, 2013.

N V Chawla, K W Bowyer, L O Hall, and W P Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

D Cianci, N Hartemink, and A Ibáñez-Justicia. Modelling the potential spatial distribution of mosquito species using three different techniques. *International Journal of Health Geographics*, 14(1):10, 2015.

A Conte, M Goffredo, C Ippoliti, and R Meiswinkel. Influence of biotic and abiotic factors on the distribution and abundance of Culicoides imicola and the Obsoletus Complex in Italy. *Veterinary Parasitology*, 150(4):333–344, 2007.

C Crisci, B Ghattas, and G Perera. A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, 240:113–122, 2012.

A C Cuéllar, L J Kjær, C Kirkeby, H Skovgard, S A Nielsen, A Stockmarr, G Andersson, A Lindstrom, J Chirico, R Lühken, et al. Spatial and temporal variation in the abundance of culicoides biting midges (diptera: Ceratopogonidae) in nine european countries. *Parasites & Vectors*, 11(1):112, 2018.

D R Cutler, T C Edwards, K H Beard, A Cutler, K T Hess, J Gibson, and J J Lawler. Random forests for classification in ecology. *Ecology*, 88(11): 2783–2792, 2007.

C De Liberato, F Farina, A Magliano, P Rombola, F Scholl, V Spallucci, and P Scaramozzino. Biotic and abiotic factors influencing distribution and abundance of Culicoides obsoletus group (Diptera: Ceratopogonidae) in central Italy. *Journal of Medical Entomology*, 47(3):313–318, 2010.

E Dijkstra, I J K Van der Ven, D R Hölzel, P A Van Rijn, and R Meiswinkel. Letter to the editor: Culicoides chiopterus as a potential vector of bluetongue virus in Europe. *Veterinary Record*, 162(13):422–422, 2008.

E Ducheyne, M A M Chueca, J Lucientes, C Calvete, R Estrada, G-J Boender, E Goossens, E M De Clercq, and G Hendrickx. Abundance modelling of invasive and indigenous culicoides species in spain. *Geospatial Health*, 8(1):241–254, 2013.

EFSA. *Bluetongue: control, surveillance and safe movement of animals*, volume 15. Panel on Animal Health and Welfare, 2017.

J Elith, J R Leathwick, and T Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697, 2009.

Alba Estrada and Beatriz Arroyo. Occurrence vs abundance models: Differences between species with varying aggregation patterns. *Biological conservation*, 152:37–45, 2012.

Jeffrey S Evans, Melanie A Murphy, Zachary A Holden, and Samuel A Cushman. Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology*, pages 139–159. Springer, 2011.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

A Fielding. *Machine learning methods for ecological applications*. Springer Science & Business Media, 1999.

C Foxi, M Pinna, V S I Monteys, and G Delrio. An updated checklist of the culicoides latreille (diptera: Ceratopogonidae) of sardinia (italy), and seasonality in proven and potential vectors for bluetongue virus (btv). *Proceedings of the Entomological Society of Washington*, 113(4): 403–416, 2011.

M Ganter. Bluetongue disease — Global overview and future risks. *Small Ruminant Research*, 118(1):79–85, 2014.

M González, S López, B A Mullens, T Baldet, and A Goldarazena. A survey of culicoides developmental sites on a farm in northern spain, with a brief review of immature habitats of european species. *Veterinary Parasitology*, 191(1-2):81–93, 2013.

S Gubbins, S Carpenter, M Baylis, J L N Wood, and P S Mellor. Assessing the risk of bluetongue to UK livestock: uncertainty and sensitivity analyses of a temperature-dependent model for the basic reproduction number. *Journal of the Royal Society Interface*, 5(20):363–371, 2008.

N Hartemink, S O Vanwambeke, B V Purse, M Gilbert, and H Van Dyck. Towards a resource-based habitat approach for spatial modelling of vector-borne disease risks. *Biological Reviews*, 90(4):1151–1162, 2015.

T Hastie, R Tibshirani, and J Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

S I Hay, A J Tatem, A J Graham, S J Goetz, and D J Rogers. Global environmental data for mapping infectious disease distribution. *Advances in Parasitology*, 62:37–77, 2006.

Tomislav Hengl, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

M A Hill. The life-cycle and habits of culicoides impunctatus goetghebuer and culicoides obsoletus meigen, together with some observations on the life-cycle of culicoides odibilis austen, culicoides pallidicornis kieffer, culicoides cubitalis edwards and culicoides chiopterus meigen. *Annals of Tropical Medicine & Parasitology*, 41(1): 55–115, 1947.

B Hoffmann, B Bauer, C Bauer, H-J Bätza, M Beer, P-H Clausen, M Geier, J M Gethmann, E Kiel, and G Liebisch. Monitoring of putative vectors of bluetongue virus serotype 8, Germany. *Emerging Infectious Diseases*, 15(9):1481, 2009.

A Ibañez-Justicia and D Cianci. Modelling the spatial distribution of the nuisance mosquito species Anopheles plumbeus (Diptera: Culicidae) in the Netherlands. *Parasites & Vectors*, 8(1):258, 2015.

N Japkowicz and S Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

Alberto Jiménez-Valverde, Francisco Diniz, Eduardo B de Azevedo, and Paulo AV Borges. Species distribution models do not account for abundance: the case of arthropods on terceira island. In *Annales Zoologici Fennici*, pages 451–464. JSTOR, 2009.

C Kaufmann, I C Steinmann, D Hegglin, F Schaffner, and A Mathis. Spatio-temporal occurrence of Culicoides biting midges in the climatic

regions of Switzerland, along with large scale species identification by MALDI-TOF mass spectrometry. *Parasites & Vectors*, 5(1):246, 2012.

D S Kettle. The bionomics and control of Culicoides and Leptoconops (Diptera, Ceratopogonidae= Heleidae). *Annual Review of Entomology*, 7(1):401–418, 1962.

M Kuhn and K Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

M Larska, L Lechowski, M Grochowska, and J F Żmudziński. Detection of the Schmallenberg virus in nulliparous Culicoides obsoletus/scoticus complex and c. punctatus—the possibility of transovarial virus transmission in the midge population and of a new vector. *Veterinary Microbiology*, 166(3-4):467–473, 2013.

A Liaw, M Wiener, et al. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.

N Lunardon, G Menardi, and N Torelli. Rose: A package for binary imbalanced learning. *R Journal*, 6(1), 2014.

MP Martin, TG Orton, E Lacarce, J Meersmans, NPA Saby, JB Paroissien, C Jolivet, L Boulonne, and D Arrouays. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. *Geoderma*, 223:97–107, 2014.

JANA M McPHERSON, Walter Jetz, and David J Rogers. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of applied ecology*, 41(5): 811–823, 2004.

H Mehlhorn, V Walldorf, S Klimpel, G Schaub, E Kiel, R Focke, G Liebisch, A Liebisch, D Werner, C Bauer, et al. Bluetongue disease in Germany (2007–2008): monitoring of entomological aspects. *Parasitology Research*, 105(2):313, 2009.

R Meiswinkel, T Baldet, R De Deken, W Takken, J-C Delécolle, and P S Mellor. The 2006 outbreak of bluetongue in northern Europe — the entomological perspective. *Preventive Veterinary Medicine*, 87(1-2): 55–63, 2008.

P S Mellor and G Prrzous. Observations on breeding sites and light-trap collections of Culicoides during an outbreak of bluetongue in Cyprus. *Bulletin of Entomological Research*, 69(2):229–234, 1979.

P S Mellor, J Boorman, and M Baylis. Culicoides biting midges: their role as arbovirus vectors. *Annual Review of Entomology*, 45(1):307–340, 2000.

P S Mellor, S Carpenter, L Harrup, M Baylis, and P PC Mertens. Bluetongue in Europe and the Mediterranean Basin: history of occurrence prior to 2006. *Preventive Veterinary Medicine*, 87(1-2):4–20, 2008.

Andy Mitchell and Michael Minami. *The ESRI guide to GIS analysis: geographic patterns & relationships*, volume 1. ESRI, Inc., 1999.

B A Mullens, A C Gerry, T J Lysyk, and E T Schmidtmann. Environmental effects on vector competence and virogenesis of bluetongue virus in Culicoides: interpreting laboratory data in a field context. *Veterinaria Italiana*, 40(3):160–166, 2004.

S A Nielsen, B O Nielsen, and J Chirico. Monitoring of biting midges (Diptera: Ceratopogonidae: Culicoides Latreille) on farms in Sweden during the emergence of the 2008 epidemic of bluetongue. *Parasitology Research*, 106(5):1197–1203, 2010.

OIE. *Terrestrial animal health code 2017. Volume 1*. World Organisation for Animal Health, twentieth edition, 2011. ISBN 978-92-9044-825-9. https://www.oie.int/doc/ged/D10905.PDF [Online accessed 12-June-2018].

R C Prati, G E Batista, and M C Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. In *Mexican International Conference on Artificial Intelligence*, pages 312–321. Springer, 2004.

C Probst, J M Gethmann, H Kampen, D Werner, and F J Conraths. A comparison of four light traps for collecting Culicoides biting midges. *Parasitology Research*, 114(12):4717–4724, 2015.

B V Purse, A J Tatem, S Caracappa, D J Rogers, P S Mellor, M Baylis, and A TORINA. Modelling the distributions of Culicoides bluetongue virus vectors in Sicily in relation to satellite-derived climate variables. *Medical and Veterinary Entomology*, 18:1–1–2, 2004.

B V Purse, P S Mellor, and M Baylis. Bluetongue in the Mediterranean: prediction of risk in space and time. *Frontis*, 9:125–136, 2005.

B V Purse, B J J McCormick, P S Mellor, M Baylis, J Boorman, D Borras, I Burgu, R Capela, S Caracappa, and F Collantes. Incriminating bluetongue virus vectors with climate envelope models. *Journal of Applied Ecology*, 44(6):1231–1242, 2007.

B V Purse, H E Brown, L Harrup, P P Mertens, and D J Rogers. Invasion of bluetongue and other orbivirus infections into Europe: the role of biological and climatic processes. *Revue Scientifique et Technique-Office International des Epizooties*, 27(2):427–442, 2008.

B V Purse, D Falconer, M J Sullivan, S Carpenter, P S Mellor, S B Piertney, A J Mordue, S Albon, G J Gunn, and A Blackwell. Impacts of climate, host and landscape factors on Culicoides species in Scotland. *Medical and Veterinary Entomology*, 26(2):168–177, 2012.

B V Purse, S Carpenter, G J Venter, G Bellis, and B A Mullens. Bionomics of temperate and tropical Culicoides midges: knowledge gaps and consequences for transmission of Culicoides-borne viruses. *Annual Review of Entomology*, 60:373–392, 2015.

T Rigot, A Conte, M Goffredo, E Ducheyne, G Hendrickx, and M Gilbert. Predicting the spatio-temporal distribution of culicoides imicola in sardinia using a discrete-time population model. *Parasites & Vectors*, 5 (1):270, 2012.

D J Rogers, S I Hay, and M J Packer. Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine & Parasitology*, 90(3): 225–241, 1996.

J Rushton and N Lyons. Economic impact of bluetongue: a review of the effects on production. *Veterinaria Italiana*, 51(4):401–406, 2015.

C Saegerman, D Berkvens, and P S Mellor. Bluetongue epidemiology in the European Union. *Emerging Infectious Diseases*, 14(4):539, 2008.

G Savini, M Goffredo, F Monaco, A Di Gennaro, M A Cafiero, L Baldi, P De Santis, R Meiswinkel, and V Caporale. Bluetongue virus isolations from midges belonging to the Obsoletus complex (Culicoides, Diptera: Ceratopogonidae) in Italy. *Veterinary Record*, 157(5):133, 2005.

J P W Scharlemann, D Benz, S I Hay, B V Purse, A J Tatem, G R W Wint, and D J Rogers. Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PloS One*, 3 (1):e1408, 2008.

C Schulz, E Bréard, C Sailleau, M Jenckel, C Viarouge, D Vitour, M Palmarini, M Gallois, D Höper, and B Hoffmann. Bluetongue virus serotype 27: detection and characterization of two novel variants in Corsica, France. *Journal of General Virology*, 97(9):2073–2083, 2016.

J M Schwenkenbecher, A J Mordue, and S B Piertney. Phylogenetic analysis indicates that Culicoides dewulfi should not be considered part of the Culicoides obsoletus complex. *Bulletin of Entomological Research*, 99(4):371–375, 2009.

A Sperlova and D Zendulkova. Bluetongue: a review. *Veterinarni Medicina*, 56(9):430–452, 2011.

S Steinke, R Lühken, C Balczun, and E Kiel. Emergence of culicoides obsoletus group species from farm-associated habitats in germany. *Medical and Veterinary Entomology*, 30(2):174–184, 2016.

A J Tatem, M Baylis, P S Mellor, B V Purse, R Capela, I Pena, and D J Rogers. Prediction of bluetongue vector distribution in europe and north Africa using satellite imagery. *Veterinary Microbiology*, 97(1-2): 13–29, 2003.

J-F Toussaint, C Sailleau, J Mast, P Houdart, G Czaplicki, L Demeestere, F VandenBussche, W Van Dessel, N Goris, and E Bréard. Bluetongue in Belgium, 2006. *Emerging Infectious Diseases*, 13(4):614, 2007.

J Van Doninck, B De Baets, J Peters, G Hendrickx, E Ducheyne, and N E C Verhoest. Modelling the spatial distribution of Culicoides imicola: climatic versus remote sensing data. *Remote Sensing*, 6(7):6604–6619, 2014.

J VanDerWal, L P Shoo, C N Johnson, and S E Williams. Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. *The American Naturalist*, 174(2):282–291, 2009.

R Venail, T Balenghien, H Guis, A Tran, M-L Setier-Rio, J-C Delécolle, B Mathieu, C Cetre-Sossah, D Martinez, and J Languille. Assessing diversity and abundance of vector populations at a national scale: example of Culicoides surveillance in France after bluetongue virus emergence. In *Arthropods as vectors of emerging diseases*, pages 77–102. Springer, 2012.

G J Venter, K Labuschagne, K G Hermanides, S N B Boikanyo, D M Majatladi, and L Morey. Comparison of the efficiency of five suction light traps under field conditions in South Africa for the collection of Culicoides species. *Veterinary Parasitology*, 166(3-4):299–307, 2009.

V Versteirt, T Balenghien, W Tack, and W Wint. A first estimation of culicoides imicola and culicoides obsoletus/culicoides scoticus seasonality and abundance in europe. *EFSA Supporting Publications*, 14(2), 2017.

G M Weiss and F Provost. The effect of class distribution on classifier learning: an empirical study. *Rutgers Univ*, 2001.

A Wilson and P Mellor. Bluetongue in Europe: vectors, epidemiology and climate change. *Parasitology Research*, 103(1):69–77, 2008.

A Wilson, S Carpenter, J Gloster, and P Mellor. Re-emergence of bluetongue in northern Europe in 2007. *Veterinary Record*, 161(14): 487, 2007.

E J Wittmann and M Baylis. Climate change: effects on Culicoides-transmitted viruses and implications for the UK. *The Veterinary Journal*, 160(2):107–117, 2000.

E J Wittmann, P S Mellor, and M Baylis. Using climate data to map the potential distribution of Culicoides imicola (Diptera: Ceratopogonidae) in Europe. *Revue Scientifique et Technique-Office International des Epizooties*, 20(3):731–740, 2001.

Lianjun Zhang, Jeffrey H Gove, and Linda S Heath. Spatial residual analysis of six modeling techniques. *Ecological Modelling*, 186(2): 154–177, 2005.

S Zientara and J M Sánchez-Vizcaíno. Control of bluetongue in Europe. *Veterinary Microbiology*, 165(1-2):33–37, 2013.